Albert Ali Salah
Javier Ruiz-del-Solar
Çetin Meriçli
Pierre-Yves Oudeyer (Eds.)

LNCS 7559

Human Behavior Understanding

Third International Workshop, HBU 2012
Vilamoura, Portugal, October 2012
Proceedings

Springer

Lecture Notes in Computer Science 7559

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

Albert Ali Salah Javier Ruiz-del-Solar
Çetin Meriçli Pierre-Yves Oudeyer (Eds.)

Human Behavior Understanding

Third International Workshop, HBU 2012
Vilamoura, Portugal, October 7, 2012
Proceedings

Springer

Volume Editors

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

Javier Ruiz-del-Solar
Universidad de Chile, Department of Electric Engineering
Av. Tupper 2007, Santiago, Chile
E-mail: jruizd@ing.uchile.cl

Çetin Meriçli
Carnegie Mellon University, Computer Science Department
Pittsburgh, PA 15213, USA
E-mail: cetin@cmu.edu

Pierre-Yves Oudeyer
FLOWERS Research Team, INRIA Bordeaux Sud-Ouest
33405 Talence, Cedex, France
E-mail: pierre-yves.oudeyer@inria.fr

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-34013-0 e-ISBN 978-3-642-34014-7
DOI 10.1007/978-3-642-34014-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012948301

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

Research domains for which an understanding of human behavior is a crucial need (e.g., robotics, human-computer interaction, affective computing, and social signal processing) rely on advanced pattern recognition techniques to automatically interpret complex behavioral patterns generated when humans interact with machines or with each other. This is a challenging problem, where many issues are still open, including the joint modeling of behavioral cues taking place on different time scales, the inherent uncertainty of machine-detectable evidences of human behavior, the mutual influence of people involved in interactions, the presence of long-term dependencies in observations extracted from human behavior, and the important role of dynamics in human behavior understanding. Implementing these methods on robotic platforms introduces further constraints on processing resources, tracking over time, model building, and generalization.

The Third Workshop on Human Behavior Understanding (HBU), organized as a satellite to IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012), gathered researchers dealing with the problem of modeling human behavior under its multiple facets (expression of emotions, display of relational attitudes, performance of individual or joint actions, imitation, etc.), with particular attention to implications in robotics, including additional resource and robustness constraints of robotic platforms, social aspects of human-robot interaction, and developmental approaches to robotics.

The workshop featured three invited talks by François Brémond (INRIA, France), Erol Şahin (METU, Turkey), and Oussama Khatib (Stanford University, USA).

François Brémond, in his talk entitled "Scene Understanding and Assisted Living," described scene understanding, which requires five levels of generic computer vision functionality of detection, localization, tracking, recognition, and understanding. Scene understanding systems go beyond the detection of visual features such as corners, edges, and moving regions to extract information related to the physical world that is meaningful for human operators. The aim is to achieve more robust, resilient, and adaptable computer vision functionalities by endowing them with a cognitive faculty: the ability to learn, adapt, weigh alternative solutions, and develop new strategies for analysis and interpretation. Brémond also discussed how scene understanding can be applied to home care monitoring.

In his talk on "Affordances and Concepts", Erol Şahin reviewed Gibson's popular notion of *affordance* through its different, sometimes contradictory, interpretations in fields ranging from human-computer interaction to autonomous robotics, to develop a formalization of affordances for its use at different levels of autonomous robot control. Using this formalization as a framework, he exposed methods on how robots could automatically learn to perceive affordances

in their environments, use learned affordance relations to ground symbolic planning mechanisms in the continuous sensory-motor experiences of the robot, and link them with concepts represented by verbs and nouns to communicate with humans. He concluded by pointing out future directions in this line of research, briefly discussing social affordances observed in human-robot interactions.

In the field of robotics, the motivation to emulate human movement has been driven by the desire to endow robots, humanoids in particular, with human-like movement properties. In his talk entitled "Robots and the Human", Oussama Khatib discussed the connection between humans and robots in terms of development of accurate models of the kinematics, dynamics, and actuation of human musculoskeletal systems, building full-body human motion simulations, performing motion reconstruction from captured data, as well as analysis and characterization of human movement. These developments, which are proving extremely valuable in human biomechanics, are providing new avenues for exploring human motion – with exciting prospects for novel clinical therapies, athletic training, character animation, and human performance improvement.

This proceedings volume contains the papers presented at the workshop and a summarizing paper. We received 31 submissions in total, and 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 thank the PAL project of INRIA, BAP 6531 project of Boğaziçi University, and the EUCogIII network for their financial support.

October 2012

Albert Ali Salah
Javier Ruiz-del-Solar
Çetin Meriçli
Pierre-Yves Oudeyer

Organization

Conference Co-chairs

Albert Ali Salah	Boğaziçi University, Turkey
Javier Ruiz-del-Solar	Universidad de Chile, Chile
Çetin Meriçli	Carnegie Mellon University, USA
Pierre-Yves Oudeyer	INRIA, France

Technical Program Committee

Levent Akın	Boğaziçi University, Turkey
Brenna Argall	Northwestern University, USA
Kai Arras	Albert-Ludwigs-Universität Freiburg, Germany
Sven Behnke	University of Bonn, Germany
Tony Belpaeme	University of Plymouth, UK
François Brémond	INRIA, France
Ginevra Castellano	University of Birmingham, UK
Kerstin Dautenhahn	University of Hertfordshire, UK
Özlem Durmaz İncel	Boğaziçi University, Turkey
Vanessa Evers	University of Twente, The Netherlands
Ian Fasel	University of Arizona, USA
Michael Goodrich	Brigham Young University, USA
Verena Hafner	Humboldt-University Berlin, Germany
Luca Iocchi	Rome University, Italy
Hatice Köse	Istanbul Technical University, Turkey
Ben Kröse	Univ. of Amsterdam and Hogeschool van Amsterdam, The Netherlands
Manuel Lopes	INRIA, France
Tekin Meriçli	Boğaziçi University, Turkey
Yukie Nagai	Osaka University, Japan
Mark Neerincx	Delft Univ. of Technology, The Netherlands
Catherine Pelachaud	Télécom ParisTech, France
Dennis Reidsma	University of Twente, The Netherlands
Erol Şahin	Middle East Technical University, Turkey
Komei Sugiura	NICT, Japan
Leila Takayama	Willow Garage, USA
Adriana Tapus	ENSTA ParisTech, France
Tijn van der Zant	University of Gröningen, The Netherlands
Manuela Veloso	Carnegie Mellon University, USA
Rodrigo Verschae	AMTC, Chile
Juan Wachs	Purdue University, USA
Zeynep Yücel	ATR, Japan

Additional Reviewers

Thomas Cederborg
Barış Evrim Demiröz
Jonathan Grizou
Olivier Mangin
Sao Mai Nguyen
Ronan Sicre

Table of Contents

Imitation and Learning from Demonstration

Human Behavior Understanding for Robotics

Albert Ali Salah[1], Javier Ruiz-del-Solar[2], Çetin Meriçli[3], and Pierre-Yves Oudeyer[4]

[1] Boğaziçi University, Department of Computer Engineering, Istanbul, Turkey
salah@boun.edu.tr
[2] Universidad de Chile, Department of Electric Engineering Av. Tupper 2007, Santiago, Chile
jruizd@ing.uchile.cl
[3] Carnegie Mellon University, Computer Science Department Pittsburgh, PA 15213, USA
cetin@cmu.edu
[4] FLOWERS Research Team, INRIA Bordeaux Sud-Ouest 33405 Talence, Cedex, France
pierre-yves.oudeyer@inria.fr

Abstract. Human behavior is complex, but structured along individual and social lines. Robotic systems interacting with people in uncontrolled environments need capabilities to correctly interpret, predict and respond to human behaviors. This paper discusses the scientific, technological and application challenges that arise from the mutual interaction of robotics and computational human behavior understanding. We supply a short survey of the area to provide a contextual framework and describe the most recent research in this area.

1 Introduction

Personal robots are predicted to arrive in homes and everyday life in the coming decades, and assist humans physically, socially, and/or cognitively. They are expected to become an integral part of the lives of people with physical or cognitive disabilities, for example, allowing the elderly or the handicapped to maintain a comfortable and autonomous life in their homes for a prolonged period of time. Furthermore, with a drastic paradigm shift in the industrial robotics, robots are also becoming closer to humans in the factories, where we observe a shift towards robots that can be intuitively and dynamically re-programmed by workers, and work jointly with them to achieve manufacturing and maintenance tasks.

Nevertheless, considering that the robots becoming so ubiquitous would result in their operating in uncontrolled environments and interacting with non-expert users, several challenging issues need to be addressed. One of these issues is human behavior understanding: in order to act in a useful, relevant, and socially acceptable manner, robots will need to understand the behavior of humans at various levels of abstractions (ranging from identifying the current action of the

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 1–16, 2012.

human to identifying goals in the discussion of two humans) and at various time scales (ranging from milliseconds to minutes and days).

A large body of work exists in the field of computational human-behavior understanding, and the International Workshop of Human Behavior Understanding, previously organized with a focus on pattern recognition and ambient intelligence, brings together scientific and technological responses to some of the challenges in this field [56,57]. While some of the proposed methods can be readily re-used for robots, novel scientific and technological challenges arise when one considers achieving human behavior understanding in the context of human-robot interaction:

- First, humans who interact with a social robot behave in ways that differ significantly from natural human-human interaction, and there is an associated new repertoire of behaviors and contextual interpretations. Thus, it is paramount to design techniques that understand human behavior specifically in the **context** of human-robot interaction.
- Second, and in a related manner, interaction with an intelligent system (be it a robot, or any artificial or ambient intelligence system) in the loop can produce dynamical evolution of human behavior, where new semiotic conventions can emerge [53]. New dynamic conventions (for example, through linguistic alignment) can be negotiated between a particular robot and a particular human, and a corresponding **dynamic update** of human behavior understanding is needed.
- Third, what makes robots specific as compared for example to classical intelligent ambient systems is that they typically have a rich repertoire of motor behaviors and actions. To be useful, relevant and socially acceptable, they need to act properly. This implies that techniques for human behavior understanding need to provide **internal representations** that are compatible and reusable by the robot's action system.

A second key challenge is the capability of robots to adapt to and learn from humans. Each human user may typically have its own preferences and habits, which a robot needs to infer. The interaction between learning and human behavior understanding can be expressed in two complementary directions:

- Robots need to be capable of learning dynamically how to interpret, and thus understand human multi-modal behavior. This includes for example learning the meaning of new linguistic constructs used by a human [18], learning to interpret the emotional state of particular users from para-linguistic or non-verbal behavior [34,58,38], characterizing properties of the interaction [44] or learning to guess the intention, and potentially the combinatorial structure of goals [39] of a human based on its overt behavior [1].
- Robots also need to be capable of learning new tasks or refining existing tasks through interaction with humans, for example using imitation learning or learning by demonstration [59,9,4,42]. This heavily involves the capacity for decoding linguistic and non-linguistic cues [34,58,38], feedback and guidance provided by humans, as well as inferring reusable primitives in human

behavior [39]. Thomaz and Breazeal [66] have for example shown that prior studies of how humans use social cues to teach can be transferred into highly useful mechanisms used by a robot to learn from humans. Such a study, related to the problem of how non-expert humans can teach new words to a robot, is presented in this volume [18].

Given that human behavior understanding in general needs to be at least partially learnt, and that learning new tasks from humans require human behavior understanding, a long-term challenge for research is to study what mechanisms can allow the joint developmental and potentially simultaneous learning of feedback/guidance/cueing models and new task models (see for example [35]).

At the same time, robotics offers stimulating opportunities for improving human behavior understanding, and especially to allow a deeper analysis of the semantics and structure of human behavior. Indeed, it is now widely known that the human action system mediates the understanding of other people's actions, in particular through the mirror neurons system [19]. Humans tend to interpret the meaning and the structure of other's behaviors in terms of their own action repertoire, which acts as a strong helping prior for this complex inference problem. Robots are also embodied and have an action repertoire, which can be similarly used to decode and interpret human behavior. For example, in this volume, Schillaci et al. show how generative action forward and inverse models of previously learnt motor primitives can be used to recognize ambiguous human movements, or to infer the target of a movement [61]. Mangin and Oudeyer show how biases on action representations can not only allow to infer the underlying combinatorial structure of complex movements demonstrated by humans, but also can be used to reproduce them [39].

In the next sections, we deal with the major contact points of human behavior understanding and robotics. Section 2 is a brief overview of systems for sensing human behavior, including pervasive systems, action and activity recognition. Section 3 discusses the social and affective aspects of human behavior from a robotics standpoint. Section 4 focuses on human-robot interaction, and Section 5 describes recent issues in imitation and learning from demonstration. Before concluding, we review a few relevant application areas briefly in Section 6 to show the practical implications of this line of research.

2 Sensing Human Behavior

The first task of a robot interacting with humans in uncontrolled environments is to sense the location of the interacting parties, as well as to recognize the relevant actions and activities. Since a lot of information can be gained by analyzing the context of interaction, multiple pattern recognition tasks are overlapped for this challenge.

2.1 Pervasive Systems

Pervasive systems describe a paradigm in which computational elements enhance interaction and intelligence of environments and objects of interaction

in a person's daily life. While many sensors are used to collect data to guide these systems, visual sensors provide perhaps the richest data over short periods [54]. François Brémond describes five levels of computer vision functionality for understanding a scene: those of detection, localization, tracking, recognition and understanding. Especially for localization, vision based sensors provide the highest accuracy for acceptable convenience levels. While recently popularized RGB-D camera technologies provide fast and accurate body tracking, most RGB-D cameras operate in limited ranges, and only under controlled illumination conditions. For mobile robots, the use of these cameras have proven to be very useful, as face-to-face interaction with humans usually occurs over small distances. The depth camera based approaches also seem to help with the high computational demands of the traditional vision-based solutions.

Cameras installed in a smart environment are typically static, configured to cover a maximal area of interest. It is possible to use multiple cameras to deal with problems of occlusion and view angles that may not be adequate at any given situation, but multi-camera systems require more complex algorithms to integrate information coming from different cameras, and are subsequently more difficult to deploy. In [14] a low-cost silhouette-based pose representation is obtained from multiple cameras and fused for action recognition.

It is obvious that installing sensors on a robot is fundamentally different than deploying the sensors on a smart environment. While the former provides a certain flexibility, it is limited by resource constraints of the robot. A promising approach to overcome some of these limits is the combination of sensors in a smart environment with the sensors on the robot. In [23], a Bayesian framework is described where a ceiling mounted camera is used for detection and tracking of people in conjunction with a laser range finder located on a mobile robot.

2.2 Action and Activity Recognition

Understanding human action mostly boils down to finding good representations of the sensed primitives. The chosen representation should be rich enough to differentiate between the action classes targeted by the application, but often it is not chosen to be much richer than that. The reason for this is purely pragmatic; more powerful representations require correspondingly complex training procedures, more training samples for learning, and longer computation time during operation. Consequently, the human body, for instance, is often represented by a graph structure made up of nodes representing landmark points on the body, and edges that connect these nodes in a fixed topology. Refinement on such a representation may be achieved by adding more landmarks (i.e. nodes) to the body parts being modeled.

In approaches where interest points do not necessarily correspond to known landmarks, space-time corners and similar 'salient' points are detected and used for learning spatio-temporal representations of actions [33]. In the present volume, Çeliktutan et al. propose an approach to solve the point set matching problem for establishing the correspondence between an action, represented by

interest points, to a template [13]. In [14], silhouettes are used for action recognition. The action template in this case is a bag of key poses representing the action in a temporal sequence.

3 Social and Affective Signals

Action recognition literature mostly focuses on simple actions, performed by a single actor [48]. A broad class of actions, however, are social in nature, and require either detailed analysis of multiple actors performing in tandem, or the distinction of very fine cues that can easily change the meaning of an action semantically. For instance, it takes a very small cue, like the creasing of the eye corners to change the meaning of a smile. Social signal processing arose from the need of intelligent systems interacting with humans to interpret and reproduce social signals, and to increase the sensitivity of the computer (or of the robot) to the interacting person's emotional and mental state [7,55]. Social signals are communicative or informative signals or cues "that directly or indirectly provide information about 'social facts': social interactions, social emotions, social attitudes, evaluations and stances, social relations, and social identities." [47].

3.1 Multimodal Analysis of Social Signals

Humans convey social information in many different ways. Facial expressions, posture, gait, body and hand gesture, speech, vocal prosody, and nonverbal cues like turn-taking behavior can all contain information relevant for interactions. Not all these signals are consciously or cognitively produced. In the present volume, Vincze et al. discuss problems that arise when people provide a certain information in a vague or approximate way, as well as the case where detectable cognitive qualities are associated with conveying information, like hesitation or hastiness [68]. An important point we made in the Introduction section of this paper is that semiotic conventions need to be established between a robot and a human in communication. While vagueness can arise because of an information gap, it can also be a device to leave open the goals designated in the communicated message. What would, for instance, be the benefit of employing vagueness when communicating with a robot? It can very well be to set up a situation where the robot decides on the correct level of abstraction or a most plausible resolution of the vague reference by examining other information available to it. This is a flexibility people have in human-human communication, and would eventually require in human-robot communication.

In natural interactions, humans also emit signals that have no real counterpart for robots. Research into human behavior understanding creates methods of analyzing these signals, which will open up new response patterns for robotic systems. In [38], an algorithm is described to determine a laughter index from visual input. This research is part of the EU-ICT FET Project ILHAIRE, which is aimed at endowing machines with automated detection, analysis, and synthesis of laughter. The authors use psychophysical descriptions of the laughter process

and propose a set of features including shoulder and body movement energy and periodicity. Obviously, a better understanding of the features that lead to accurate detection of laughter will also help us build systems that can synthesize realistic instances of laughter.

3.2 Perception of Affect

Emotions are important modifiers of human behavior, serving to enrich the response palette, but also allowing faster and contextualized decisions to help the human function better. Part of the importance of emotions also comes from the fact that humans are quite adept at recognizing emotional displays in others, and this forms the backbone of a social existence. In fact, this capability is so strong that humans easily attribute affect even to technological artifacts, as the well known Heider-Simmel study has demonstrated with simple moving geometric shapes [22]. In [58], Hylozoic Soil, a responsive architectural geotextile environment, is used to induce affective responses in viewers. Basic emotions like anger, sadness and happiness can be conveyed with simple movements of these dynamic structures. The authors also establish that there are gender differences in the perception of these affective movements [58]. These studies confirm that social interaction between humans and robots cannot ignore the affective dimension.

Movement is rarely used for automatic affect analysis of humans. In face to face communication, robots can observe the facial expressions of the interacting humans, as well as analyze the voice for affective signals. These are the most typically used modalities for affect analysis. In the present volume, Lim and Okuno show that a robot can also use the gait of a person to determine affective states [34]. In their approach, speed, intensity, irregularity, and extent features are extracted from the gait and speech of persons to determine affective states like happiness, sadness, anger, and fear. The advantage of using gait is that the face may not be available to a robot at all times, and the movement and resolution of the face may make emotion recognition difficult.

Ziemke and Lowe characterize emotion as (a) being closely connected to embodied cognition, (b) grounded in homeostatic bodily regulation, and (c) a powerful and useful organizational principle for modulation of behavioral and cognitive mechanisms [70]. Their focus is on maintaining emotion as an integral part of the internal environment of a robot, and as they admit, the role of emotion in social interactions is not addressed in their work, but they do note that the interplay of internal (i.e. individual) and external (i.e. social) aspects of emotion is still not very well known [3]. Robotic platforms can be excellent experimental tools for probing into these relatively unexplored areas.

4 Human-Robot Interaction

One of the long term ambitious goals of robotics research is to have robots capable of seamlessly integrating themselves in our daily environments. Therefore, recognizing, interpreting, and reasoning about the human behavior is a critical

skill for a robot that co-inhabits the human environments and interacts with humans on a regular basis. Particularly difficult challenges in human behavior understanding from the robotics point of view are the necessity to perform the processing using the limited computational resources on board, and using the sensors that can be mounted on a robotic platform.

4.1 Interacting with Robots

In general, the human-robot interaction (HRI) research can be divided into two main categories:

- **Human-centered HRI** investigates issues like the design and usability of proper interaction interfaces, robot platforms, and behaviors through extensive user studies.
- **Robot-centered HRI** focuses on algorithms, engineering innovations, and other computational approaches that would improve the overall performance of the interaction.

Although there is no clear distinction, the majority of the research on synthesizing behaviors, facial expressions and whole body gestures, and the development of proper interaction media fall into the human-centered HRI branch, especially from the validation point of view, while perceiving and interpreting behaviors, recognizing speech, and interactive learning applications fall into the robot-centered HRI branch.

A good example of the first approach is [28] in this volume, which reports the use case development for an outdoor robotic tour guide. In this work, abstractions of human behaviors appropriate for robot tour guides were developed. These abstractions form the basis of implemented robotic behaviors, which are then assessed in the real application scenario, where the robot meets visitors in a fairly unconstrained manner.

4.2 Closing the Interaction Loop

In the present volume, Fischer and Saunder investigate how people's initial expectations from an interaction, and their increasing experience and acquaintance with the robot over prolonged interaction sessions affect the way people tend to interact with robots [18]. Speech-based interaction has been heavily studied over the past decade. Grounding spatial commands given using unrestricted natural language for commanding a robot to navigate in the environment and manipulate objects have been studied in [65,24,29].

Humans also use gaze and gestures heavily to narrow down the uncertainties about the context when conversing verbally. Especially, forming joint attention through modeling the gaze of a human can be very useful in human-robot collaboration scenarios or when a human teacher teaches tasks or concepts involving the objects in the environment [69,63]. In [69], object saliency is used in conjunction with head pose estimates to allow a humanoid robot to determine the

visual focus of attention of the interacting human, while in [63] a fixed mapping between head pose directions and gaze target directions was not assumed, and models are investigated that perform a dynamic (temporal) mapping implicitly accounting for varying body/shoulder orientations of a person over time, as well as unsupervised adaptation.

Closing the interaction loop requires robots that behave closer to humans, and have more exploratory behavior than currently allowed for. An important concept related to the exploration capabilities of the robot is the notion of *"Symbiotic Autonomy"*. Accepting the fact that the robot has physical and cognitive limitations, and assuming the robot is also aware of some of its limitations, symbiotic autonomy advocates the benefits of engaging with the humans in the environment in a symbiotic relationship so that the robot does tasks for people, and asks people for help whenever its capabilities fall short of dealing with a certain situation [50]. Human interaction with the objective of asking for help raises new challenges like how and where to find humans who would likely provide help [51,52], and if there are more than one human present in the scene, whom to approach, as well as how to approach. Especially for the latter case, the ability to infer the intent of people as well as their predicted movement trajectories can drastically improve the way the robot interacts with the humans, and hence, the quality of the help it receives.

5 Imitation and Learning from Demonstration

Imitation is a process of paramount importance in both human-human and human-robot interaction. It is used for diverse functions, ranging from interaction regulation and social bonding to learning new knowledge and new competencies from others. In the recent years, imitation has been highly explored in various robotics contexts: its role for natural, intuitive and usable human-robot social interaction [46], robot learning of new tasks from demonstration [6,4], and its origins and functions in the course of epigenesis in developmental robotics [2,27,5]. Imitation learning in particular poses fundamental and challenging scientific problems [45], related to what, when and who to imitate, and it may be achieved at various levels of abstractions. Lopes et al. [36] describe three main levels of abstraction in imitation, which are respectively addressed by three chapters in this book: Mimicking behavior and trajectory-level imitation [44], imitation mediated by the action system and motor primitives [61], and imitation of goals and intentions [39].

The first level of abstraction in imitation is **mimicking**, where imitation consists of directly trying to reproduce the observed movements without an attempt to infer their underlying structure or goals [36]. A large amount of methods and approaches have been developed within this approach, and more particularly in the context of imitation learning where many researchers have studied how machine learning regression techniques could be used to reproduce smooth and generalizing trajectories out of sets of noisy human demonstrations (e.g. [6,11,20,12]). Besides robot learning of new skills by trajectory-level imitation, it

is also highly useful that robots be able to detect when two humans, or a human and a robot, are imitating each other [44]. Michelet et al. propose a combined computer vision and machine learning approach, which targets automatic identification of imitative interaction among humans [44]. A useful feature of this approach for easy deployment in the real-world is that the approach is capable of analyzing fine aspects of movements without the need to identify and track human skeletons.

The second level of abstraction in imitation leverages the **mediation of the action repertoire**, and consists in first interpreting observed behavior in terms of one's own repertoire of motor primitives, which are then re-used to generate the imitation [36]. Such approaches have been gaining popularity in robot learning recently, in particular because such an approach allows to reduce the dimensionality of movement and behavior representations meaningfully, which in turn often allows for better robustness and generalization [41,26]. A central question within these approaches, both in biology and robotics, is to understand how these motor primitives form initially. Some works have explored various learning techniques that allow to automatically infer and learn motion primitives from observation of human behaviors, e.g. [60,32,31,40]. In the present volume, Schillaci et al. study a complementary question [61]: once motor primitives have been learnt - in this case with the help of an annotated database -, how can they be used to recognize human actions and disambiguate potential targets? Using such a motor primitive representation, in the form of paired forward and inverse models, is highly useful, since it can allow direct reproduction of the observed behavior.

The third level of abstraction in imitation is **goal imitation** [36]. Here, the imitator tries to infer the intention, or the goal of the observed behavior, and then tries to reproduce this goal, possibly with different means (for example with a different motor policy). Mathematically, this amounts to learning the hidden cost function that the observed behavior may try to maximize, and then using this cost function to define a surrogate optimization/learning problem which the imitator has to solve. From a theoretical point of view, these approaches have been studied in two fields, optimal feedback control [49] and inverse reinforcement learning (IRL) [1], respectively. In recent years, they have been applied to imitation learning in robotics, where they have been shown to be powerful for generalization and robust to environment change at the same time [1,67]. For example, Abbeel and Ng[1] showed how autonomous helicopters could learn to achieve acrobatic flights better than professional human demonstrators using this approach. Lopes et al. [37] showed how active learning techniques could be used to increase the efficiency of IRL. In the present volume, Mangin and Oudeyer explore a frontier of these approaches [39]: how can a robot learn the combinatorial structure of the hidden goals underlying demonstrated behaviors? Previous approaches assumed that observed behavior corresponds to a single hidden cost function/goal. On the contrary, Mangin and Oudeyer consider the case when the demonstrator has a repertoire of hidden goals and only produces behaviors which concurrently target several goals. The proposed approach relies

on establishing a bridge between inverse feedback control techniques and dictionary learning techniques [25]. Like [61], [39] infers a motor representation of observed behaviors of a demonstrator, which allows the learning system to both recognize and reproduce behavior with adequate generalization.

6 Applications

In the future, we envision robots that not only assist humans in domestic environments, but also interact with them in public spaces and factories. These robotic applications will require proper social interactions to be maintained between robots and humans. We will move from the current situation in which robots carry out certain tasks without any interaction with humans (e.g. automotive factory) to situations in which humans and robots will co-work, and subsequently robots becoming co-workers and co-inhabitants [21], carrying out tasks of increasing complexity that require understanding the behavior of humans. Even the simple task of cooperatively carrying a table by a human and a robot [64], requires the synchronization of the individual's movements, which can only be achieved if the other's behavior is correctly analyzed in real-time with the correct resolution and level of abstraction.

In this section, we give some application examples to make the requirements more concrete.

6.1 Socially Assistive Robotics

One of the envisioned applications of robotics is assisting specific human populations, such as children, elderly people, and patients. These are tasks that require specific expertise in relatively restricted domains, embodiment, and most importantly, a social aspect that makes robots preferable to automated systems that are less suited to display and interpret social and affective signals. Socially assistive robotics defines the robot's goal to be the creation of "close and effective interaction with a human user for the purpose of giving assistance and achieving measurable progress in convalescence, rehabilitation, learning, etc." [17]. Some related applications are robots as exercise coaches, evaluating the moves of the interacting humans [16], and guiding robots providing context-dependent information to people [28].

In these applications human behavior understanding will be crucial for interpreting the human needs and requirements, but also for understanding the mood and for taking actions to manage it appropriately. Understanding human moods and needs requires having some basic functionality. One of them is understanding the visual focus of attention of humans while interacting with robots. This is addressed in this volume [63].

Applications in which robots interact with humans have increased largely in the last years thanks to the development of the Microsoft Kinect sensor. The sensor's ability of obtain 3D-images at a low cost, and the availability of libraries with functionalities such as human body segmentation, have boosted

the development of HRI applications. Most of these applications are related with entertainment, although they can be expanded to education. However, the use of infrared lighting makes the sensor being surface dependent (e.g. on black surfaces the reflection is very limited), and not appropriate to be used outdoors. Therefore, complementary sensors need to be used in applications with those constraints.

6.2 Playful Interactions

The work of Cynthia Breazeal and others has established that people interacting with robots will treat the robot as a social entity [8,10]. Consequently, robots have the potential to be much more than elaborate toys in children's games. In social games of children, interactions are not pre-determined, but emerge through mutual interaction. The ideal game partner is thus one that adapts to a game scenario, and one that can assume one of many different roles, each as coherent as possible in the social and affective displays that belong to the designated role. The contribution of human behavior understanding to this kind of a scenario would be the detailed analysis of gaming roles to create the coherent role models, as well as real-time observation of the playing partners to determine which mode should be selected and put into action.

A less ambitious, but worthy goal is to use robots as mediators in playful interactions. A very important research direction is for instance the work with autistic children, who may shun social contact in the form provided by their peers, but may come to like what a social robot has to offer. An example is the work of Michaud and Théberge-Turmel, who used robust robotic toys in play experiments with children to obtain promising results [43]. The AURORA project is an important initiative in this area with the aim of to encouraging autistic children "to become engaged in a variety of different interactions important to human social behavior" [15]. Another good example of toy robots for interacting with children is the Keepon robot, which is capable of conveying limited emotion and attention, promoting social playful interaction [30].

The basic idea that underlies these applications is that play is a fundamental activity in learning social interactions. While human behavior understanding has been used in gaming scenarios in the design phase, to specify interaction scenarios, real-time behavior analysis is only recently being integrated into games [62].

7 Conclusions

In this introductory paper of the 3rd International Workshop on Human Behavior Understanding, our primary aim was to articulate the points of contact between robotics and human behavior understanding. It is clear that progress in the latter will have direct bearing on the design and implementation of robots that have social skills and interact with humans in more natural ways. The proper approach to do this is not mere imitation of the human behavior, but goes through a deeper understanding of the abstract processes leading to particular behavior and ways of interaction, so as to let the counterparts emerge in

the interacting robots. Obviously, a lot of basic skills must be in place before this can be achieved.

The second important point is what robotics has to offer to human behavior understanding, especially in terms of new scientific questions it poses. Since robots need to act in an embodied manner, it is essential that human-behavior understanding capabilities provided to/learned by robots are adapted to allow leveraging this understanding (e.g. the representations) to act appropriately. Purely functional representations may not be sufficient, and robotics is an excellent testbed for this; if the correct abstraction is not achieved, transferring behavior patterns to the robot will not be successful.

A final point is that the presence of the robots causes changes in the behavior of humans. It is important to understand what kind of new social situations are created by putting robots with social capabilities, and social roles, in a natural environment. As the skill palette of robots grows, and they start reading and responding to social and affective displays of humans, these mutual relationships will be increasingly complex, and will require more thorough analysis.

Acknowledgments. This work is supported by INRIA project PAL, Boğaziçi University project BAP-6531, EUCogIII, and ERC EXPLORERS 240007.

References

1. Abbeel, P., Ng, A.Y.: Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the 21st International Conference on Machine Learning (ICML 2004), pp. 1–8 (2004)
2. Andry, P., Gaussier, P., Nadel, J., Hirsbrunner, B.: From sensori-motor development to low-level imitation. Adaptive Behavior 12, 117–138 (2004)
3. Arbib, M.A., Fellous, J.M.: Emotions: from brain to robot. Trends in Cognitive Sciences 8(12), 554–561 (2004)
4. Argall, B.D., Chernova, S., Veloso, M., Browning, B.: A survey of robot learning from demonstration. Robotics and Autonomous Systems 57(5), 469–483 (2009)
5. Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., Yoshida, C.: Cognitive developmental robotics: A survey. IEEE Trans. Autonomous Mental Development 1(1) (2009)
6. Billard, A., Calinon, S., Dillmann, R., Schaal, S.: Survey: Robot programming by demonstration. In: Handbook of Robotics, ch. 59 (2008)
7. Breazeal, C.: Emotion and sociable humanoid robots. International Journal of Human-Computer Studies 59(1-2), 119–155 (2003)
8. Breazeal, C.: Toward sociable robots. Robotics and Autonomous Systems 42(3), 167–175 (2003)
9. Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., Blumberg, B.: Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. Artificial Life 11(1-2), 31–62 (2005)
10. Brooks, A.G., Gray, J., Hoffman, G., Lockerd, A., Lee, H., Breazeal, C.: Robot's play: interactive games with sociable machines. Computers in Entertainment (CIE) 2(3), 1–10 (2004)

11. Calinon, S., Guenter, F., Billard, A.: On learning, representing and generalizing a task in a humanoid robot. IEEE Transactions on Systems, Man and Cybernetics, Part B 37(2), 286–298 (2007)
12. Cederborg, T., Li, M., Baranes, A., Oudeyer, P.-Y.: Incremental local inline gaussian mixture regression for imitation learning of multiple tasks. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan (2010)
13. Çeliktutan, O., Wolf, C., Sankur, B., Lombardi, E.: Real-Time Exact Graph Matching with Application in Human Action Recognition. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 17–28. Springer, Heidelberg (2012)
14. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: An Efficient Approach for Multi-view Human Action Recognition Based on Bag-of-Key-Poses. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 29–40. Springer, Heidelberg (2012)
15. Dautenhahn, K., Werry, I.: Towards interactive robots in autism therapy: Background, motivation and challenges. Pragmatics & Cognition 12(1), 1–35 (2004)
16. Fasola, J., Matarić, M.J.: Robot exercise instructor: A socially assistive robot system to monitor and encourage physical exercise for the elderly. In: 19th IEEE International Symposium in Robot and Human Interactive Communication, Viareggio, Italy, pp. 416–421 (September 2010)
17. Feil-Seifer, D., Matarić, M.J.: Defining socially assistive robotics. In: 9th International Conference on Rehabilitation Robotics, ICORR 2005, pp. 465–468. IEEE (2005)
18. Fischer, K., Saunders, J.: Between Initial Expectations and Acquaintance: Interacting with a Developing Robot. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 125–133. Springer, Heidelberg (2012)
19. Gobbini, M.I., Koralek, A.C., Bryan, R.E., Montgomery, K.J., Haxby, J.V.: Two takes on the social brain: A comparison of theory of mind tasks. Journal of Cognitive Neuroscience 19(11), 1803–1814 (2007)
20. Grollman, D.H., Jenkins, O.C.: Sparse incremental learning for interactive robot control policy estimation. In: International Conference on Robotics and Automation (ICRA 2008), pp. 3315–3320 (May 2008)
21. Guizzo, E., Deyle, T.: Robotics trends for 2012 (the future is robots). IEEE Robot. Automat. Mag. 19(1), 119–123 (2012)
22. Heider, F., Simmel, M.: An experimental study of apparent behavior. The American Journal of Psychology 57(2), 243–259 (1944)
23. Hu, N., Englebienne, G., Kröse, B.: Bayesian Fusion of Ceiling Mounted Camera and Laser Range Finder on a Mobile Robot for People Detection and Localization. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 41–51. Springer, Heidelberg (2012)
24. Huang, A.S., Tellex, S., Bachrach, A., Kollar, T., Roy, D., Roy, N.: Natural language command of an autonomous micro-air vehicle. In: Int. Conf. on Intelligent Robots and Systems (IROS), Taipei, Taiwan (October 2010)
25. Jenatton, R., Mairal, J., Obozinski, G., Bach, F.: Proximal Methods for Hierarchical Sparse Coding. Journal of Machine Learning Research 12, 2297–2334 (2011), http://hal.inria.fr/inria-00516723
26. Jenkins, O.C., Matarić, M.J., Weber, S.: Primitive-based movement classification for humanoid imitation. In: IEEE International Conference on Humanoid Robots, Humanoids 2000 (2000)

27. Kaplan, F., Oudeyer, P.-Y.: The progress-drive hypothesis: an interpretation of early imitation. In: Dautenhahn, K., Nehaniv, C. (eds.) Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions. Cambridge University Press (2007)
28. Karreman, D.E., Evers, V., van Dijk, E.M.A.G.: Contextual Analysis of Human Non-verbal Guide Behaviors to Inform the Development of FROG, the Fun Robotic Outdoor Guide. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 113–124. Springer, Heidelberg (2012)
29. Kollar, T., Tellex, S., Roy, D., Roy, N.: Toward understanding natural language directions. In: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010, pp. 259–266. IEEE Press, Piscataway (2010), http://dl.acm.org/citation.cfm?id=1734454.1734553
30. Kozima, H., Michalowski, M.P., Nakagawa, C.: Keepon: A playful robot for research, therapy, and entertainment. International Journal of Social Robotics 1(1), 3–18 (2009)
31. Krüger, V., Herzog, D., Baby, S., Ude, A., Kragic, D.: Learning actions from observations. IEEE Robot. Automat. Mag. 17(2), 30–43 (2010)
32. Kulić, D., Nakamura, Y.: Incremental Learning of Full Body Motion Primitives. In: Sigaud, O., Peters, J. (eds.) From Motor Learning to Interaction Learning in Robots. SCI, vol. 264, pp. 383–406. Springer, Heidelberg (2010)
33. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64(2), 107–123 (2005)
34. Lim, A., Okuno, H.G.: Using Speech Data to Recognize Emotion in Human Gait. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 52–64. Springer, Heidelberg (2012)
35. Lopes, M., Oudeyer, P.-Y.: Active learning and intrinsically motivated exploration in robots: Advances and challenges (guest editorial). IEEE Transactions on Autonomous Mental Development 2(2), 65–69 (2010)
36. Lopes, M., Melo, F., Montesano, L., Santos-Victor, J.: Abstraction Levels for Robotic Imitation: Overview and Computational Approaches. In: Sigaud, O., Peters, J. (eds.) From Motor Learning to Interaction Learning in Robots. SCI, vol. 264, pp. 313–355. Springer, Heidelberg (2010)
37. Lopes, M., Melo, F., Montesano, L.: Active Learning for Reward Estimation in Inverse Reinforcement Learning. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part II. LNCS, vol. 5782, pp. 31–46. Springer, Heidelberg (2009)
38. Mancini, M., Varni, G., Glowinski, D., Volpe, G.: Computing and Evaluating the Body Laughter Index. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 90–98. Springer, Heidelberg (2012)
39. Mangin, O., Oudeyer, P.-Y.: Learning the Combinatorial Structure of Demonstrated Behaviors with Inverse Feedback Control. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 135–147. Springer, Heidelberg (2012)
40. Mangin, O., Oudeyer, P.-Y.: Learning to recognize parallel combinations of human motion primitives with linguistic descriptions using non-negative matrix factorization. To Appear in IEEE/RSJ International Conference on Intelligent Robots and Systems (2012)
41. Matarić, M.J.: Sensory-motor primitives as a basis for learning by imitation:linking perception to action and biology to robotics. In: Imitation in Animals and Artifacts. MIT Press (2002)

42. Meriçli, Ç., Veloso, M., Akın, H.L.: Improving biped walk stability with complementary corrective demonstration. Autonomous Robots 32(4), 419–432 (2012), http://dx.doi.org/10.1007/s10514-012-9284-1
43. Michaud, F., Théberge-Turmel, C.: Mobile robotic toys and autism. Socially Intelligent Agents, 125–132 (2002)
44. Michelet, S., Karp, K., Delaherche, E., Achard, C., Chetouani, M.: Automatic Imitation Assessment in Interaction. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 161–173. Springer, Heidelberg (2012)
45. Nehaniv, C.: Nine billion correspondence problems. In: Dautenhahn, K., Nehaniv, C. (eds.) Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions. Cambridge University Press (2007)
46. Nehaniv, C.L., Dautenhahn, K. (eds.): Imitation and social learning in robots, humans, and animals: behavioural, social and communicative dimensions. Cambridge University Press (2004)
47. Poggi, I., D'Errico, F.: Social signals: a framework in terms of goals and beliefs. Cognitive Processing (2012)
48. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
49. Ratliff, N., Bagnell, J., Zinkevich, M.: Maximum margin planning. In: Proc. 23rd Int. Conf. Machine Learning, pp. 729–736 (2006)
50. Rosenthal, S., Biswas, J., Veloso, M.: An effective personal mobile robot agent through symbiotic human-robot interaction. In: International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010), vol. 1, pp. 915–922 (May 2010)
51. Rosenthal, S., Veloso, M.M., Dey, A.K.: Acquiring accurate human responses to robots' questions. I. J. Social Robotics 4(2), 117–129 (2012)
52. Rosenthal, S., Veloso, M.M., Dey, A.K.: Is someone in this office available to help me? - proactively seeking help from spatially-situated humans. Journal of Intelligent and Robotic Systems 66(1-2), 205–221 (2012)
53. Salah, A., Schouten, B.: Semiosis and the relevance of context for the AmI environment. In: Proc. European Conf. on Computing and Philosophy (2009)
54. Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A.: Computer vision for ambient intelligence. Journal of Ambient Intelligence and Smart Environments 3(3), 187–191 (2011)
55. Salah, A.A., Pantic, M., Vinciarelli, A.: Recent developments in social signal processing. In: 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 380–385. IEEE (2011)
56. Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A.: Challenges of Human Behavior Understanding. In: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (eds.) HBU 2010. LNCS, vol. 6219, pp. 1–12. Springer, Heidelberg (2010)
57. Salah, A.A., Lepri, B., Pianesi, F., Pentland, A.: Human Behavior Understanding for Inducing Behavioral Change: Application Perspectives. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 1–15. Springer, Heidelberg (2011)
58. Samadani, A.-A., Gorbet, R., Kulić, D.: Gender Differences in the Perception of Affective Movements. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 65–76. Springer, Heidelberg (2012)
59. Schaal, S., Ijspeert, A., Billard, A.: Computational approaches to motor learning by imitation. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358(1431), 537–547 (2003)

60. Schaal, S., Peters, J., Nakanishi, J., Ijspeert, A.: Learning movement primitives. In: International Symposium on Robotics Research, ISRR 2003 (2003)
61. Schillaci, G., Lara, B., Hafner, V.: Internal Simulations for Behaviour Selection and Recognition. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 148–160. Springer, Heidelberg (2012)
62. Schouten, B.A.M., Tieben, R., van de Ven, A., Schouten, D.W.: Human Behavior Analysis in Ambient Gaming and Playful Interaction. In: Salah, A.A., Gevers, T. (eds.) Computer Analysis of Human Behavior, pp. 387–403. Springer-Verlag London Limited (2011)
63. Sheikhi, S., Odobez, J.-M.: Recognizing the Visual Focus of Attention for Human Robot Interaction. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 99–112. Springer, Heidelberg (2012)
64. Stuckler, J., Holz, D., Behnke, S.: Robocup@home: Demonstrating everyday manipulation skills in robocup@home. IEEE Robotics Automation Magazine 19(2), 34–42 (2012)
65. Tellex, S., Kollar, T., Dickerson, S., Walter, M.R., Banerjee, A.G., Teller, S., Roy, N.: Understanding natural language commands for robotic navigation and mobile manipulation. In: Proceedings of the National Conference on Artificial Intelligence (AAAI) (August 2011)
66. Thomaz, A.L., Breazeal, C.: Teachable robots: Understanding human teaching behavior to build more effective robot learners. Artificial Intelligence Journal 172, 716–737 (2008)
67. Verma, D., Rao, R.: Goal-based imitation as probabilistic inference over graphical models. In: Advances in NIPS 18 (2006)
68. Vincze, L., Poggi, I., D'Errico, F.: Vagueness and Dreams. Analysis of Body Signals in Vague Dream Telling. In: Salah, A.A., Ruiz-del Solar, J., Meriçli, Ç., Oudeyer, P.-Y. (eds.) HBU 2012. LNCS, vol. 7559, pp. 77–89. Springer, Heidelberg (2012)
69. Yücel, Z., Salah, A.A., Meriçli, Ç., Meriçli, T.: Joint visual attention modeling for naturally interacting robotic agents. In: 24th International Symposium on Computer and Information Sciences, ISCIS 2009 (2009)
70. Ziemke, T., Lowe, R.: On the role of emotion in embodied cognitive architectures: From organisms to robots. Cognitive Computation 1(1), 104–117 (2009)

Real-Time Exact Graph Matching with Application in Human Action Recognition

Oya Çeliktutan[1,2], Christian Wolf[2], Bülent Sankur[1], and Eric Lombardi[2]

[1] Electrical and Electronics Engineering,
Boğaziçi University, Istanbul, Turkey
[2] Université de Lyon, CNRS,
INSA-Lyon, LIRIS, UMR CNRS 5205, F-69621, France

Abstract. Graph matching is one of the principal methods to formulate the correspondence between two set of points in computer vision and pattern recognition. However, most formulations are based on the minimization of a difficult energy function which is known to be NP-hard. Traditional methods solve the minimization problem approximately. In this paper, we show that an efficient solution can be obtained by exactly solving an approximated problem instead of approximately solving the original problem. We derive an exact minimization algorithm and successfully apply it to action recognition in videos. In this context, we take advantage of special properties of the time domain, in particular causality and the linear order of time, and propose a novel spatio-temporal graphical structure.

Keywords: Spatio-temporal graph, Hyper-graph matching, Action recognition.

1 Introduction

In many applications involving the recognition of complex visual patterns, for instance recognition of object classes or actions in video scenes, salient local features collected on sparse set of points provide a compact yet rich representation, for classification or matching. This approach can be robust, e.g. against occlusion and bypasses the tedious segmentation task. The resulting representation is inherently structural and is therefore difficult to use in a statistical learning framework without sacrificing all or a part of the spatial or spatio-temporal relationships. In fact, the ensemble of local features is often converted into a numerical representation, discarding all or most of the structural information in the process. A typical example is the bag-of-words (BoW) formalism, originally developed for image classification [1]. However, graphs (and hyper-graphs) form a natural description of this type of data.

In the context of human action recognition, a graph can effectively represent the relationship between low-level features such as spatio-temporal interest points, descriptors, human body parts etc. In [2], a number of interest points is structured into a graph per frame. Distance between a scene frame-graph and a set of prototypes are fed to Hidden Markov Models (HMMs) for classification. In [3], the relationship between different entities, e.g. spatio-temporal descriptors and spin-images, is modeled by a graph embedded into a common Euclidean space. In [4], the nodes correspond to the five body

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 17–28, 2012.

parts and the energy function is penalized with the constraints on the human body configuration.

In this work, we concentrate on hyper-graph matching and point set matching, where the nodes of the graph(s) encode both position and description of spatio-temporal interest points, and the neighborhood relationship is derived from proximity information. Matching corresponds to finding an action model point set in a (usually larger) scene point set. Up to our knowledge, prior work on space-time graph matching can be summed up by a few recent papers. In [5], matching is done via temporally ordered local feature-graphs where each graph models spatial configuration of the features in a small temporal segment. Graphs are built from adjacency relationships of space-time tubes produced from oversegmenting the test video in [6], and from proximity by thresholding distances in space and time in [7]. These methods resort to off-the-shelf spectral methods or slightly modified versions of them. In contrast, our proposed method takes advantage of some properties of the 3D space in which the data is embedded to devise an exact algorithm.

There are alternative approaches taking into account space-time 3D geometry: In [8], the interest points are divided into clusters where each cluster is modeled by its relative spatial position as well as the distribution of the appearance and position of interest points. In [9], the correlation of spatio-temporal (ST) patterns is measured and ST correlograms are constructed. Pairwise spatio-temporal relations are introduced in [10], based on a set of rules, and this information is transformed into 3D histograms. In [11], interest points, optical flow and image segmentation are mixed, and classification is done with multiple search trees. In [12], a parts-based model integrates spatio-temporal configuration, appearance, and human-object interactions. Finally, in [13], a branch-and-cut algorithm searches the best scoring subgraph over a learned spatio-temporal graph for each action class.

The linear nature of the time dimension is frequently used to devise methods based on sequence alignment. In [14], a chain graph model exploits a priori knowledge of the nature and semantics of the relationship between different variables. More examples are trajectory matching with Gabor filters [15], learning salient state transitions by HMMs [16], and modeling the evaluation of silhouettes over time [17].

Our proposed algorithm is related to sequence alignment in that it exploits temporal information and its linear nature in a similar way. However, we do not perform simple sequence alignment. The novelty of our approach is that we use a full-fledged hyper-graph model with all its rich structural information stored in its nodes, embedded in space-time, and in its hyper-edges built from proximity information. The derived minimization algorithm is capable of dealing with classical energy functions including unary, binary and ternary terms, which makes it possible to include scale invariant potentials, as the formulations in [18,19,20] and others. Once the graph representation of a given video sequence is obtained, action recognition problem boils down to searching for the closest prototype graph in the graph-space. Overview of our approach is illustrated in Figure 1.

Techniques for graph matching and for point set matching have been studied intensively in pattern recognition. While the graph isomorphism problem can be calculated in polynomial time, it is widely known that exact subgraph matching is NP-complete

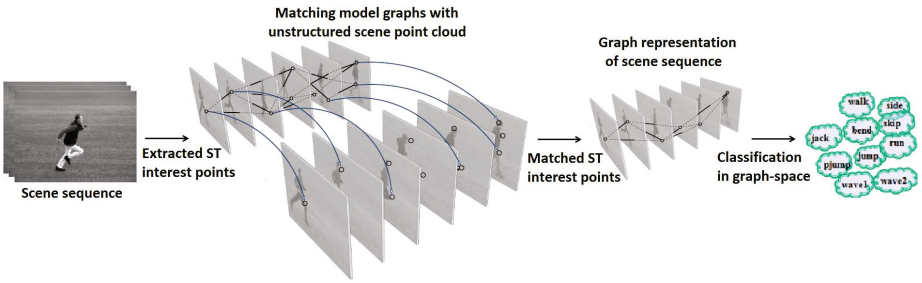


Fig. 1. Overview of the proposed algorithm for action recognition

[20], as is subgraph isomorphism [21]. In the context of object recognition, a method which approximates the graph, which in turn enables computation of the exact solution in polynomial time has been proposed in [22]: a k-tree is built randomly from the spatial interest points on an object, which allows for the application of the classical junction tree algorithm [23]. Spectral methods such as in [24] relax the binary assignment problem into a continuous one and show that the solution for the continuous problem is the principal eigenvector of the constraints matrix. The solution of the original problem is calculated by thresholding the solution of the continuous problem, which is an approximation — the discrete optimum is not necessarily related to a continuous solution. In [19], this is extended to hyper-graphs and the Eigenproblem is solved efficiently with an iterative algorithm. In [25], a convex-concave programming approach is employed on a least-squares problem of the permutation matrices. Several methods decompose the original problem into sub problems which are solved with different optimization tools like graph cuts [20,26]. In [27], a multi-label graph cuts minimizer is extended to 2D problems by alternating between labels and nodes. In [28], a candidate graph structure is created and the problem is formulated as a multiple coloring problem on a layered structure. A solution for the resulting integer quadratic programming problem is advanced in [29], the problem is extended to relationships of general order (> 3) and solved with random walks. Finally, in a related paper dynamic programming and graph algorithms [30] are described.

The contributions in this paper are two-fold:

- A theoretical result stating that for the data embedded in space-time, the exact solution to the point set matching problem with hyper-graphs can be calculated in complexity exponential on a small number, which becomes bounded when the hyper-graph is structured using proximity relationships.
- A practical solution to the action recognition problem in videos applying the proposed algorithm to graphs designed with a special structure. This allows calculating matches with computational complexity, which grows linearly in the number of model nodes and linearly in the number of scene nodes.

The paper is organized as follows: Section 2 formulates the graph matching problem and discusses related work on the problem. Section 3 discusses the special properties of

the space in which our data are embedded. In Section 4, we propose a special structure of our model graphs and derive an algorithm which further reduces the computational complexity of the matching algorithm. Section 5 describes the experiments and Section 6 finally concludes.

2 Problem Formulation

In this paper, we formulate the problem as a particular case of the general correspondence problem between two point sets. The objective is to assign points from the model set to points in the scene set, such that some geometrical invariance is satisfied. We solve the problem through a global energy minimization which takes into account a hyper-graph[1] constructed from the model point set. The M points of the model are organized as a hyper-graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes (corresponding to the points) and $\mathcal{E}$ is the set of edges. From now on we will abusively call hyper-graphs "graphs" and hyper-edges "edges". The edges $\mathcal{E}$ in our graph connect sets of three nodes, thus triangles.

While our method requires the data in the model video to be structured into a graph, this is not necessarily so for the data in the scene video. While structural information on the scene data *can* be integrated easily into our formulation, which allows adding structural terms into the minimization framework, and thus results in a classical graph-matching problem. Our formulation is thus more general but can also deal with graph matching.

Each point i of the two sets (model and scene) is also assigned a position $p_i = [\, p_i^{<x>}\ p_i^{<y>}\ p_i^{\langle t \rangle}\,]^T$ and a feature vector f_i describing the appearance of a local space-time region around this point. When necessary, we will distinguish between model and scene values by the superscripts $^{\langle m \rangle}$ and $^{\langle s \rangle}$: $p_i^{\langle m \rangle}, f_i^{\langle m \rangle}, p_i^{\langle s \rangle}, f_i^{\langle s \rangle}$ etc. Note that symbols in superscripts enclosed in angle brackets $\langle . \rangle$ are not numerical indices, they are mere symbols indicating a category.

Each node i of the model graph is assigned a discrete variable $x_i,\ i = 1..M$, which represents the mapping from the ith model node to some scene node, and can take values from $\{1 \dots S\}$, where S is the number of scene nodes. The whole set of variables x_i is also abbreviated as x. A solution of the problem is given through the values of the x_i, where a value of $x_i = j$ is interpreted as model node i being assigned to scene node j. To handle occlusions, an additional dummy value ϵ is admitted, which semantically means that no assignment has been found for the given variable.

Each combination of assignments x evaluates to an energy value using an energy function $E(x)$. In principle, the energy should be lower for assignments that correspond to a realistic transformation from the model image to the scene image, and it should be high otherwise. We search for the assignments that minimize this energy.

Using pairwise edges mostly restricts geometrical coherence constraints to distance similarities, which are not invariant to scale changes. Higher order matching through hyper-graphs has been proposed in the context of object recognition [24]. Typically,

[1] A hyper-graph is a generalization of a graph, where a hyper-edge can connect any number of vertices, typically more than two [31].

hyper-edges connect 3 nodes, which allows to formulate geometrical constraints between pairs of triangles. In particular, geometrical similarity can be measured through angles, which are scale invariant. Our proposed energy function is of the following form:

$$E(x) = \lambda_1 \sum_i U(x_i) + \lambda_2 \sum_{(i,j,k)\in\mathcal{E}} D(x_i, x_j, x_k) \tag{1}$$

where U is a data attached term taking into account feature distances, D is the space-time geometric distortion between two triangles and λ_1 and λ_2 are weighting parameters. For convenience, all dependencies on all values over which we do not optimize have been omitted. U is defined as the Euclidean distance between the appearance features of assigned points, taking into account a penalty W^P for the dummy assignment:

$$U(x_i) = \begin{cases} W^p & \text{if } x_i = \epsilon, \\ ||{f_i}^{\langle m \rangle} - f_{x_i}^{\langle s \rangle}|| & \text{else.} \end{cases} \tag{2}$$

Since our data is embedded in space-time, angles are projections from 3D+t to 2D, thus include a temporal component not related to scale changes induced by zooming. We therefore split the geometry term D into a temporal distortion term D^t and a spatial geometric distortion term D^g:

$$D(x_i, x_j, x_k) = D^t(x_i, x_j, x_k) + \lambda_3 D^g(x_i, x_j, x_k) \tag{3}$$

where the temporal distortion D^t is defined as truncated time differences over two pairs of nodes of the triangle and geometric distortion D^g is defined over differences of angles.

3 Space-Time Matching

In our work, the geometric data are embedded in space-time. We assume the following commonly accepted properties of space-time to derive an efficient algorithm:

Hypothesis 1: Causality — Each point in the two sets (i.e., model and scene) lies in a 3-dimensional space : $(p_i^{\langle x \rangle}, p_i^{\langle y \rangle}, p_i^{\langle t \rangle})$. The spatial and temporal dimensions should not be treated in the same way. While objects (and humans) can undergo arbitrary geometrical transformations like translation and rotation, which is subsumed by geometrical matching invariance in our problem, human actions can normally *not* be reversed.

In a correct match, the temporal order of the points should be retained, which can be formalized as follows

$$\forall\, i, j : p_i^{\langle m \rangle \langle t \rangle} \leq p_j^{\langle m \rangle \langle t \rangle} \Leftrightarrow p_{x_i}^{\langle s \rangle \langle t \rangle} \leq p_{x_j}^{\langle s \rangle \langle t \rangle} \tag{4}$$

Let us recall that the superscript $\langle t \rangle$ stands for the time dimension, and it is not an index.

Hypothesis 2: Temporal closeness — Another reasonable assumption is that the extent of time warping between model and scene time axes must be limited. In other words, two points which are close in time must be close in both the model set and

the scene set. This property can be used to further decrease the search space during inference. Since our graph is created from proximity information (we threshold space-time distances between nodes to extract the hyper-edges), it can be formalized as

$$\forall\, i,j,k \in \mathcal{E} : |p_{x_i}^{\langle s\rangle\langle t\rangle} - p_{x_j}^{\langle s\rangle\langle t\rangle}| < T^t \;\vee\; |p_{x_j}^{\langle s\rangle\langle t\rangle} - p_{x_k}^{\langle s\rangle\langle t\rangle}| < T^t \tag{5}$$

Hypothesis 3: Unicity of time instants — We assume that time instants cannot be split or merged. In other words, all points of the same model frame should be matched to points of the same scene frame.

$$\begin{aligned} \forall i,j : (p_i^{\langle m\rangle\langle t\rangle} = p_j^{\langle m\rangle\langle t\rangle}) \Leftrightarrow (p_{x_i}^{\langle s\rangle\langle t\rangle} = p_{x_j}^{\langle s\rangle\langle t\rangle}) \wedge \\ (p_i^{\langle m\rangle\langle t\rangle} \neq p_j^{\langle m\rangle\langle t\rangle}) \Leftrightarrow \;\; (p_{x_i}^{\langle s\rangle\langle t\rangle} \neq p_{x_j}^{\langle s\rangle\langle t\rangle}) \end{aligned} \tag{6}$$

In [32], we showed that (under these hypotheses) the complexity of exactly minimizing in Eq. (1) is exponential only on the maximum number of points per frame, which is typically a small number, e.g., $1-4$. However, in practice and for general graphs it is still too high for practical usage. The next section will introduce a special structure which further decreases complexity.

4 A Special Graphical Structure

Recall that classical methods use approximate solutions since exact minimization of formulations such as in Eq. (1) is infeasible. In this work, we advocate an alternative and perhaps better idea, which is to approximate the problem — the graphical structure in this case — and to solve the new problem exactly. This is particular appealing in point matching problems where the structure of the graph is less related to the description of the object, but rather to the constraints of the matching process.

We propose to structure the model points as follows:

- We keep a single point in each model frame by choosing the most salient one, i.e. the ones with the highest confidence of the interest point detector. However, no restrictions are applied to the scene frames, which may contain an arbitrary number of points.
- Each model point i is connected to its two immediate predecessors $i-1$ and $i-2$ as well as to its two immediate successors $i+1$ and $i+2$.

This creates a planar graph with triangular structure, as illustrated in Figure 2. The general case of the energy function (1) can be simplified in this case. The neighborhood system can be described in a very simple way using the index of the nodes of the graph, similar to the dependency graph of a second order Markov chain:

$$E(x) = \sum_{i=1}^{M} U(x_i) + \sum_{i=3}^{M} D(x_i, x_{i-1}, x_{i-2}). \tag{7}$$

The general recursive formula of the inference algorithm can be derived as

$$\alpha_i(x_{i-1}, x_{i-2}) = \min_{x_i} \Big[\, U(x_i) + D(x_i, x_{i-1}, x_{i-2}) + \; \alpha_{i+1}(x_i, x_{i-1}) \,\Big] \tag{8}$$

with the initialization

$$\alpha_M(x_{M-1}, x_{M-2}) = \min_{x_M} \left[U(x_M) + D(x_M, x_{M-1}, x_{M-2})\right]. \tag{9}$$

During the calculation of the trellis, the arguments of the minima in equation (8) are stored in a table $\beta_i(x_{i-1}, x_{i-2})$. Once the trellis completed, the optimal assignment can be calculated through classical backtracking:

$$\widehat{x_i} = \beta_i(x(i-1), x(i-2)), \tag{10}$$

starting from an initial search for x_1 and x_2:

$$(\widehat{x_1}, \widehat{x_2}) = \arg\min_{x_1, x_2} [U(x_1) + U(x_2) + \alpha_3(x_1, x_2)]. \tag{11}$$

The algorithm as given above is of complexity $O(M{\cdot}S^3)$: a trellis is calculated in a $M \times S \times S$ matrix, where each cell requires to iterate over S possible combinations.

Exploiting the different hypotheses on the spatio-temporal data introduced in section 2, the complexity can be decreased further:

Ad) Hypothesis 1 — taking causality constraints into account we can prune many combinations from the trellis of the optimization algorithm. In particular, if we calculate possibilities in the trellis given a certain assignment for a given variable x_i, all values for the predecessors x_{i-1} and x_{i-2} must be necessarily *before* x_i, i.e. lower.

Ad) Hypothesis 2 — similar as above, given a certain assignment for a given variable x_i, we will allow a maximum number of T^t possibilities for the values of the successors x_{i-1}, x_{i-2}, which are required to be *close*.

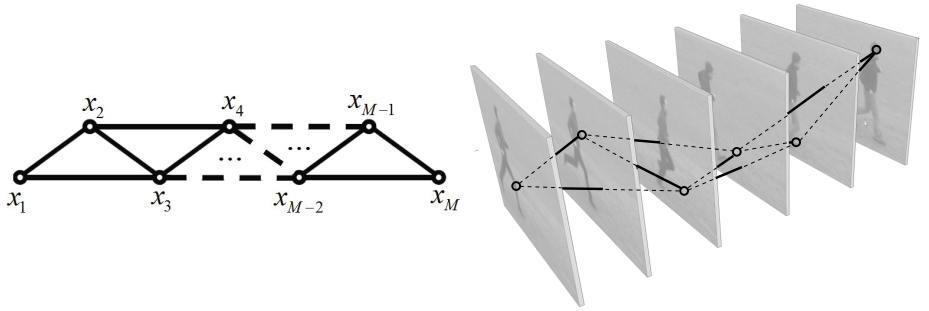
Thus, the expression in equation (8) is only calculated for combinations satisfying the following constraints:

$$\begin{array}{ccccc} |x_i - x_{i-1}| < T^t & \wedge & |x_{i-1} - x_{i-2}| < T^t & \wedge \\ x_i > x_{i-1} & \wedge & x_{i-1} > x_{i-2}. & \end{array} \tag{12}$$

These pruning measures decreases the complexity to $O(M{\cdot}S{\cdot}{T^t}^2)$, where T^t is a small constant measured in the number of frame, so the complexity is linear on the number of points in the scene: $O(M{\cdot}S)$. For example, let the number of model nodes and scene nodes be $M = 30$ and $S = 500$, respectively, we achieve 2500 fold complexity reduction when $T^t = 10$.

5 Experimental Results

We tested the proposed method on the widely used public KTH dataset [34]. It includes 25 subjects performing 6 actions (*walking*, *jogging*, *running*, *handwaving*, *handclapping* and *boxing*) recorded in four different scenarios including indoor/outdoor scenes and different camera viewpoints. Spatio-temporal interest points extracted with the 3D Harris detector [33] constitute the nodes of the proposed graphical structure. Appearance features f_i are the well known HoG/HoF extracted with the publicly available code



(a) (b)

Fig. 2. (a) A special graphical structure for the model point set designed for very low computational complexity: a second order chain. However, no requirements are imposed on the scene point set. (b) An example model graph.

Table 1. Confusion matrix without (a) and with (b) prototype selection. Respective accuracies: 86.3%, 90.6%. (B: Box, HC: Handclap, HW: Handwave, J: Jog, R: Run, W: Walk).

	B	HC	HW	J	R	W
B	**97**	3	0	0	0	0
HC	0	**100**	0	0	0	0
HW	3	16	**81**	0	0	0
J	0	0	0	**71**	29	0
R	0	0	0	25	**75**	0
W	0	0	0	3	3	**94**

(a)

	B	HC	HW	J	R	W
B	**100**	0	0	0	0	0
HC	0	**100**	0	0	0	0
HW	13	6	**81**	0	0	0
J	0	0	0	**78**	19	3
R	0	0	0	10	**85**	5
W	0	0	0	0	0	**100**

(b)

Table 2. Comparison with existing methods using the same KTH dataset protocol. (B: Box, HC: Handclap, HW: Handwave, J: Jog, R: Run, W: Walk).

Method	B	HC	HW	J	R	W	Tot.
Laptev et al. [33]	97	95	91	89	80	99	91.8
Schuldt et al. [34]	98	60	74	60	55	84	71.8
Li et al. [35]	97	94	86	**100**	83	97	**92.8**
Niebles et al. [36]	99	97	**100**	78	80	94	91.3
Our method	**100**	**100**	81	78	**85**	**100**	90.6

in [33]. As mentioned in section 4, we choose a single point per model frame based on the confidence score of the detector. All points are kept for testing videos.

The parameters are fixed as follows. The penalty parameter W_P should theoretically be higher than the average local energy of correctly assigned triangles and lower than the average local energy of incorrectly assigned triangles. We estimate it by sampling energies (without penalty) of pairs of training sequences in two settings: intra-class and inter-class, resulting in two histograms of local energies. We set $W^P = 4$ as the point of minimal Bayes error. The weighting parameters are optimized over the validation set: $\lambda_1 = 0.6$, $\lambda_2 = 0.2$, $\lambda_3 = 5$, $T^t = 10$, and $W^t = 20$.

First, we build up a model dictionary using leave-one-subject-out (LOSO) strategy. We generate several model graphs by partitioning the sequences into subsequences each containing between 20 to 30 number of frames with salient interest points. This results in approximately 2200 model graphs in total. Action classes on the unseen subjects are recognized with a nearest neighbor (NN) classifier where the distance is defined as the matching energy (1). The average recognition performance of the proposed scheme is found to be 86.3%. The main cause of this modest performance is the poor discrimination between the *jogging* and *running* classes (see Table 1a). The algorithm also suffers from *handwaving*, while significantly successful in *boxing*, *handclapping* and *walking*. We conjecture that this issue can be handled by a prototype selection algorithm.

Prototype selection — In prototype-based approaches, prototype selection plays a key role in recognition performance. Intra-variation can be large among action categories; some categories include different numbers of views or different categories can be similar, thus misleading, in the graph-space constructed. We balanced and optimized the dictionary with Sequential Floating Backward Search (SFBS), which removes irrelevant model graphs from the training set. SFBS has been successfully used as a supervised feature selection method in many previous studies [37]. Briefly, we start with a full dictionary and proceed to remove conditionally the least significant models from the set, one at a time, while checking the performance variations. Deletions which improve the performance are made permanent in this greedy search. After a number of removal steps, we reintroduce one or more of the removed ones provided they improve the performance. At each step, performance is evaluated on a validation set. We use the same data partition protocol (8/8/9) as proposed in [34]. We select 44 models as our best subset of model graphs, which increased test performance to 90.6%. As expected, the *jogging* and *running* sequences benefit the most from dictionary learning (see Table 1b).

Sample matched model and scene sequences are illustrated in Figure 3. While the first action (*handwaving*) is successfully recognized, the second one (*running*) gives an example of misclassification. Table 2 proves that our method has a comparable performance with state-of-the-art methods. We want to point out that many results have been published on the KTH database, but the protocols are not comparable for most of them, see the review in [38]. In the figure, we chose results obtained with the same data partition protocol [34].

The algorithm has been implemented in Matlab. Matching each model graph is done simultaneously with 0.02 seconds per frame, i.e. for an average scene of 30 seconds ($S = 750$) recognition takes 13.8 seconds on a CPU with 3.33GHz and 4GB RAM.

A real-time GPU implementation– A first preliminary GPU implementation allows real-time performance on standard medium end GPUs, e.g. a Nvidia GeForce GTS450. Table 3 compares run times of the CPU implementation in Matlab/C and the GPU implementation running on different GPUs with different characteristics, especially the number of calculation units. The run times are given for matching a single model graph with 30 nodes against scene blocks of different lengths. If the scene video is cut into smaller blocks of 60 frames, which is necessary for continuous video processing, real time performance can be achieved even on the low end GPU model. With these smaller

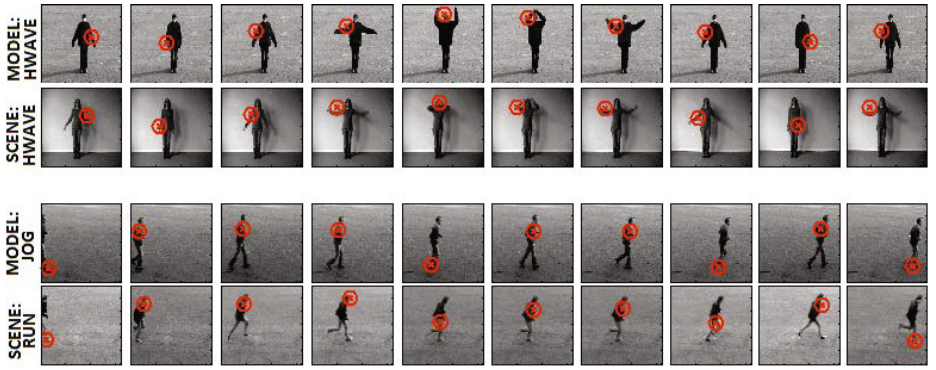


Fig. 3. Examples for matched sequences: Upper match results in correct recognition, while the lower match is misclassified

Table 3. Running times in milliseconds for two different GPUs and for 4 different scene block sizes. The last column on the right gives times per frame for matching the whole set of 44 model graphs.

Implementation	Nodes	Frames	Time (ms)	Time/fr (ms)	Time/fr (ms)
			— — A single model — —		— All 44 models —
CPU: Intel Core 2 Duo, E8600 @ 3.33Ghz, Matlab/C(mex)	754	723	13800	19.09	840
Nvidia GeForce GTS450, 192 cuda cores, 128 bit memory interface	754 60	723 55	748 4	1.03 0.07	45 3 (real time)
Nvidia GeForce GTX560, 336 cuda cores, 256 bit memory interface	754 60	723 55	405 4	0.56 0.07	25 (real time) 3 (real time)

chunks of scene data, matching all 44 graph models to a block of 60 frames (roughly 2 seconds of video) takes roughly 3ms regardless of the GPU model.

The processing time of 3ms/fr is very much lower than the limit for real time processing, which is 40ms for video acquired at 25fps. Additional processing will be required in order to treat overlapping blocks, which increases running time to 6ms/fr. The times given above also do not include interest point detection and feature extraction, but these are negligible compared to the matching requirements and can also be calculated on a GPU.

6 Conclusions and Future Work

In this paper we showed that — when the data is embedded in space-time — the exact solution to the point set matching problem with hyper-graphs can be calculated in

complexity exponential on a small number, which is bounded when the hyper-graph is structured with proximity information. As a second contribution we presented a special graphical structure which allows to perform exact matching with very low complexity, linear in the number of the model nodes and the number of scene nodes. The method has been tested on the KTH dataset where it shows competing performance with very low runtime.

Our current work concentrates on extension of graphical structure to more than one interest point per frame. This idea is formulated through "meta" graph or "frame" graph matching in which each node in the graph corresponds to a frame of the video and each frame is characterized by spatio-temporal interest points and triangles. Following this, we will use more challenging videos.

References

1. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: ICCV, vol. 2, pp. 1470–1477 (2003)
2. Borzeshi, E.Z., Piccardi, M., Xu, R.Y.D.: A discriminative prototype selection approach for graph embedding in human action recognition. In: ICCVW (2011)
3. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In: CVPR (2008)
4. Raja, K., Laptev, I., Prez, P., Oisel, L.: Joint pose estimation and action recognition in image graphs. In: ICIP (2011)
5. Gaur, U., Zhu, Y., Song, B., Roy-Chowdhury, A.: A string of feature graphs model for recognition of complex activities in natural videos. In: ICCV (2011)
6. Brendel, W., Todorovic, S.: Learning spatiotemporal graphs of human activities. In: ICPR (2011)
7. Ta, A.P., Wolf, C., Lavoue, G., Başkurt, A.: Recognizing and localizing individual activities through graph matching. In: AVSS (2010)
8. Niebles, J.C., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. In: CVPR, pp. 1–8 (2007)
9. Savarese, S., Delpozo, A., Niebles, J., Fei-Fei, L.: Spatial-temporal correlatons for unsupervised action classification. In: WMVC, Los Alamitos, CA (2008)
10. Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: ICCV (2009)
11. Mikolajczyk, K., Uemura, H.: Action recognition with appearance motion features and fast search trees. CVIU 115(3), 426–438 (2011)
12. Filipovych, R., Ribeiro, E.: Robust sequence alignment for actor-object interaction recognition: Discovering actor-object states. CVIU 115, 177–193 (2011)
13. Chen, C., Grauman, K.: Efficient activity detection with max-subgraph search. In: CVPR (2012)
14. Zhang, L., Zeng, Z., Ji, Q.: Probabilistic image modeling with an extended chain graph for human activity recognition and image segmentation. IEEE Tr. on IP (2011)
15. Dyana, A., Das, S.: Trajectory representation using gabor features for motion-based video retrieval. Pattern Recognition Letters 30(10), 877–892 (2009)
16. Cuntoor, N.P., Yegnanarayana, B., Chellappa, R.: Activity modeling using event probability sequences. IEEE Tr. on IP 17(4), 594–607 (2008)
17. Abdelkader, M.F., Abd-Almageed, W., Srivastava, A., Chellappa, R.: Silhouette-based Gesture and Action Recognition via Modeling Trajectories on Riemannian shape manifolds. CVIU 115(3), 439–455 (2010)

18. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. IJPRAI 18(3), 265–298 (2004)
19. Duchenne, O., Bach, F.R., Kweon, I.-S., Ponce, J.: A tensor-based algorithm for high-order graph matching. In: CVPR, pp. 1980–1987 (2009)
20. Torresani, L., Kolmogorov, V., Rother, C.: Feature Correspondence Via Graph Matching: Models and Global Optimization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 596–609. Springer, Heidelberg (2008)
21. Zampelli, S., Deville, Y., Solnon, C.: Solving subgraph isomorphism problems with constraint programming. Constraints (2009)
22. Caetano, T.S., Caelli, T., Schuurmans, D., Barone, D.A.C.: Graphical models and point pattern matching. IEEE Tr. on PAMI 28(10), 1646–1663 (2006)
23. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society B 50, 157–224 (1988)
24. Leordeanu, M., Hebert, M.: A spectral technique for correspondence problems using pairwise constraints. In: ICCV, Washington, DC, USA, pp. 1482–1489 (2005)
25. Zaslavskiy, M., Bach, F., Vert, J.P.: A path following algorithm for the graph matching problem. IEEE Tr. on PAMI 31(12), 2227–2242 (2009)
26. Zeng, Y., Wang, C., Wang, Y., Gu, X., Samaras, D., Paragios, N.: Dense non-rigid surface registration using high-order graph matching. In: CVPR (2010)
27. Duchenne, O., Joulin, A., Ponce, J.: A graph-matching kernel for object categorization. In: ICCV (2011)
28. Lin, L., Zeng, K., Liu, X., Zhu, S.-C.: Layered graph matching by composite cluster sampling with collaborative and competitive interactions. In: CVPR, vol. 0, pp. 1351–1358 (2009)
29. Leordeanu, M., Zanfir, A., Sminchisescu, C.: Semi-supervised learning and optimization for hypergraph matching. In: ICCV (2011)
30. Felzenszwalb, P.F., Zabih, R.: Dynamic programming and graph algorithms in computer vision. IEEE Tr. on PAMI 33(4), 721–740 (2011)
31. Zass, R., Shashua, A.: Probabilistic graph and hypergraph matching. In: CVPR (2008)
32. Çeliktutan, O., Wolf, C., Sankur, B.: Fast exact matching and correspondence with hypergraphs on spatio-temporal data. LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumiére Lyon 2/Ecole Centrale de Lyon Report No. RR-LIRIS-2012-002 (2012)
33. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR, pp. 1–8 (2008)
34. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR, pp. 32–36 (2004)
35. Li, B., Ayazoğlu, M., Mao, T., Camps, O.I., Sznaier, M.: Activity recognition using dynamic subspace angles. In: CVPR (2011)
36. Niebles, J.C., Chen, C.-W., Fei-Fei, L.: Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 392–405. Springer, Heidelberg (2010)
37. Pudil, P., Ferri, F.J., Novovicov, J., Kittler, J.: Floating search methods for feature selection with non-monotonic criterion functions. In: ICPR, pp. 279–283 (1994)
38. 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)

An Efficient Approach for Multi-view Human Action Recognition Based on Bag-of-Key-Poses

Alexandros Andre Chaaraoui[1], Pau Climent-Pérez[1], and Francisco Flórez-Revuelta[2]

[1] Department of Computing Technology, University of Alicante, Alicante, Spain
{alexandros,pcliment}@dtic.ua.es
[2] Faculty of Science, Engineering and Computing, Kingston University, Kingston upon Thames, United Kingdom
F.Florez@kingston.ac.uk

Abstract. This paper presents a novel multi-view human action recognition approach based on a bag-of-key-poses. In the case of multi-view scenarios, it is especially difficult to perform accurate action recognition that still runs at an admissible recognition speed. The presented method aims to fill this gap by combining a silhouette-based pose representation with a simple, yet effective multi-view learning approach based on *Model Fusion*. Action classification is performed through efficient sequence matching and by the comparison of successive key poses which are evaluated on both feature similarity and match relevance. Experimentation on the MuHAVi dataset shows that the method outperforms currently available recognition rates and is exceptionally robust to actor-variance. Temporal evaluation confirms the method's suitability for real-time recognition.

Keywords: human action recognition, multi-view action recognition, key pose, bag-of-key-poses, MuHAVi dataset.

1 Introduction

In human action recognition based on vision techniques, one of the first questions to address is if a single- or a multi-view based approach should be chosen. Some application scenarios (e.g. interactive robots) or field of view (FOV) restrictions (e.g. in automotive systems) can limit the number of available camera views to only one. Nevertheless, due to the reduction of costs and the increase of popularity of outdoor and indoor cameras, there are commonly several cameras installed covering the same FOV. Especially in human action recognition, one camera can be insufficient due to partial occlusions (objects like furniture could be in the way, but also other persons) and ambiguous or unfavorable viewing angles. While great effort has been made in the last ten years to improve single-view human action recognition in order to achieve high recognition rates and satisfying recognition speeds [1,2], there are still few successful multi-view methods [3]. The main reasons for this situation are: 1) the additional increase

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 29–40, 2012.

of difficulty in learning from multiple views, because the combination of multi-view data leads to a greater data variance and complex learning models; and 2) the resulting decrease of recognition speed, as at least two views need to be processed and analysed (or chosen from).

In this paper, we present a multi-view human action recognition method which targets the two aforementioned difficulties. With a simple silhouette-based pose representation, an efficient combination of multiple views is achieved by a learning approach based on *Model Fusion* using a bag-of-key-poses, similar to the bag-of-words paradigm. This leads us to a very effective method that outperforms current state-of-the-art recognition rates and still maintains its real-time suitability. At the same time, no restrictions are imposed about the points of view (POV), besides of training-testing coincidence.

The remainder of this paper is organized as follows: Section 2 summarizes recent works on human action recognition, emphasising the different ways of combining data from multiple views. Section 3 details the chosen pose representation which is used in section 4 as input for the learning process. In section 5, action recognition through sequence matching is detailed. The obtained results of recognition accuracy and speed are presented in section 6. Section 7 presents some conclusions and discussion.

2 Related Work

Regarding the type of input features used for classification in human action recognition, we can divide between *global* and *local* approaches. The former takes into account the whole image or a specific region of interest (ROI), normally defined by motion's location. In this sense, [4] presented an encoding of temporal evolution and spatial location of motion over a sequence of frames (Motion History- and Energy-Images). This has been extended by [5] to a 3D Motion History Volume in order to obtain a free-viewpoint representation. A similar goal is pursued in [6], where the temporal dimension is considered explicitly building space-time volumes based on a sequence of binary silhouettes. Space-time saliency and orientation features are used for action recognition, detection and clustering. In *local* or *sparse* representations, research interest lies in obtaining and encoding a set of points of spatial and temporal interest. Consequently, several works extended traditional salient point detectors to 3D in order to capture motion clues [7,8,9]. A different proposal is given at [10], where oriented rectangular patches combined with a histogram-based approach and Dynamic Time Warping (DTW) showed great effectiveness. For greater detail about the state-of-the-art of human action recognition we refer to [1,11].

More specifically, in multi-view scenarios, different levels of combination of visual data obtained from multiple views can be found. At the uppermost stage, information fusion is applied at *decision-level*. In other words, single-view human action recognition methods are applied individually for each view, and in the final recognition phase the test video sequence is matched with the *best-view*. Examples are found in [12,13,14], where the best view is chosen based on classification outputs as lowest distance, highest score/probability of feature matching,

number of detected features, majority voting, etc; or on criteria depending on the input data like the view with the biggest ROI or the best retrieved silhouette. Naturally, this type of multi-view action recognition is the most straightforward and simple option, and its great advantage is the ease of transition from single- to multi-view. Nevertheless, its major difficulty is the definition of an appropriate decision rule, as this strongly depends on the application scenario and the nature of the actions to recognize. Furthermore, when dealing with a greater number of views, parallel execution of single-view action recognition methods would require a distributed approach in order to maintain an acceptable recognition speed.

At the underneath level, the so-called *Model Fusion* aims to consider multiple views in the model learning phase. This can be done either implicitly, feeding the classifier with images ignoring their viewing angle [13], or explicitly, adapting the learning process to multiple views [15]. This approach can have clear performance advantages over the former, since a single learning process enables multi-view recognition. The main difficulty is to successfully change the learning scheme. Differences can be found whether or not possible POV are restricted and if a 2D or 3D model is chosen.

Another common approach is to inherently learn features proceeding from multiple views by applying *Feature Fusion*. The combination of multiple features in order to be considered as a single complex feature can be achieved, for instance, by concatenating [13] or averaging [12] the single-view features. Similarly to the information fusion applied at *decision-level*, combination of this technique with single-view action recognition methods is relatively effortless. Nonetheless, the complexity of classification may increase when adding further views.

Finally, in [16,17], *Data Fusion* is applied, since binary silhouettes obtained from multiple views are considered as 3D data before applying any feature extraction. The appeal of this level of information fusion lies in avoiding multiple information loss when generating features from raw data. However, it requires observations which can be fused and compatible feature generations.

3 Silhouette-Based Pose Representation

As introduced in section 1, our method uses a silhouette-based pose representation as input feature. Specifically, the contour points $P = \{p_1, p_2, ..., p_n\}$ of the silhouettes (extracted with a border following algorithm [18]) are used, as these preserve the spatial information but ignore the redundant interior points. Advantages of contour-based features lie in resistance to small viewpoint variations or lighting changes [19], and in the fact that morphological pre-processing steps are not needed. Concretely, we use the work by Dedeoğlu et al. presented in [20] and described shortly as follows:

1. Defining $p_i = (x_i, y_i)$, the center of mass $C_m = (x_c, y_c)$ is obtained as:

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n}, y_c = \frac{\sum_{i=1}^{n} y_i}{n} \quad . \tag{1}$$

2. A distance signal $DS = \{d_1, d_2, ..., d_n\}$ is generated, whose elements are defined as the Euclidean distance between each point and the center of mass:

$$d_i = \|C_m - p_i\|, \quad \forall i \in [1...n] \ . \tag{2}$$

3. Defining a constant size L of the distance signal and normalising its values to unit sum, scale-invariance is achieved:

$$\hat{DS}[i] = DS\left[\left\lceil i * \frac{n}{L} \right\rceil\right], \quad \forall i \in [1...L] \ , \tag{3}$$

$$\bar{DS}[i] = \frac{\hat{DS}[i]}{\sum_{i=1}^{L} \hat{DS}[i]}, \quad \forall i \in [1...L] \ . \tag{4}$$

As further detailed in section 6, this feature successfully encodes spatial information and its generation presents a very low computational cost.

4 Model Fusion of Multiple Views

Once all the video frames are processed to their silhouette-based pose representation, these samples are reduced to a representative subset of key poses, the bag-of-key-poses, which represents the most characteristic poses in time. The motivation behind using key poses is to distinguish one action from another based on a few individual poses, achieving this way a significant decrease of the problem scale by omitting redundant data.

4.1 Learning a Bag-of-Key-Poses

Figure 1 shows an overview of the learning process. Let us suppose there are M available view points and R action classes to learn. Considering a single action a, all the pose representations are reduced to K key poses per-view. This means that the pose representations of each view are considered separately so as to simplify the key pose generation process. Key poses are obtained by *K-means* clustering with Euclidean distance. Taking the cluster centers as the representatives of each group of the training data, K key poses are generated. This process is repeated for each action class ending up with a bag-of-key-poses of $R \times K \times M$ key poses.

With this definition, the bag-of-key-poses is made up of the most characteristic poses of multiple views. However, some key poses could be more discriminative than others if they can be only found in a specific action. On the other hand, key poses that represent very characteristic, but also very common poses (e.g. standing still) could provide a poor discriminative value as they are present in a wide variety of actions. For this reason, all available pose representations from all views and action classes are assigned to their *nearest neighbor* key pose of the bag-of-key-poses (based on the Euclidean distance between their features),

and the ratio of within class matches is computed to be used later as the weight w of each key pose:

$$w = \begin{cases} \frac{matches}{assignments} & \text{if } assignments > 0 \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{5}$$

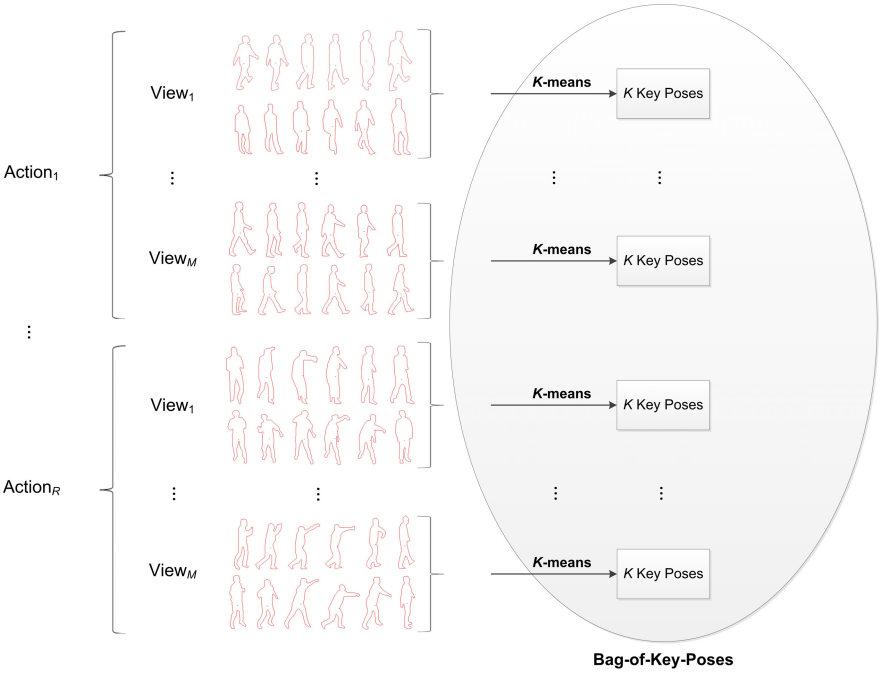


Fig. 1. Overview of the generation process of the bag-of-key-poses

4.2 Learning Sequences of Key Poses

So far, no temporal aspects have been taken into account, and frame-by-frame recognition could be handled without considering any particular motion order. Nevertheless, as video sequences of action performances are used for training, valuable information about the temporal evolution of silhouettes is available. Similarly, in an on-line recognition scenario, silhouettes would be acquired in the particular order of the subject's performance.

Consequently, the long-term temporal evolution of key poses can be modelled by finding, for all available training sequences, the *nearest neighbor* key pose kp of each video frame's pose representation. The corresponding successive *nearest neighbor* key poses compose a simplified sequence of known key poses and their temporal evolution: $S = \{kp_1, kp_2, ..., kp_t\}$. In this step, not only the temporal order of appearance of key poses is modelled, but the training data is also shifted

to the shared and defined domain of the bag-of-key-poses, what leads to noise and outlier filtering.

5 Action Sequence Recognition

At the recognition stage, the same procedure is initially performed: 1) the frames of the test video sequence are processed to their pose representation; and 2) the corresponding sequence of key poses is built by obtaining the successive *nearest neighbor* key poses. At this point, the recognition problem can be solved with a sequence matching algorithm. For this purpose, DTW has been chosen because it shows proficiency in temporal alignment of sequences that share the same temporal order, but could present inconsistent lengths, accelerations and decelerations. This is exactly our case, since action performances among humans of different age or condition can vary in speed, and the involved parts can present different motion paces.

Given two sequences of key poses $S_{train} = \{kp_1, kp_2, ..., kp_t\}$ and $S_{test} = \{kp'_1, kp'_2, ..., kp'_u\}$ we compute the DTW distance $d_{DTW}(S_{train}, S_{test})$ as:

$$d_{DTW}(S_{train}, S_{test}) = dtw\,(t, u) \quad , \tag{6}$$

$$dtw\,(i, j) = \min \left\{ \begin{array}{l} dtw\,(i-1, j)\,, \\ dtw\,(i, j-1)\,, \\ dtw\,(i-1, j-1) \end{array} \right\} + d(kp_i, kp'_j) \quad , \tag{7}$$

where $d(kp_i, kp'_j)$ is the distance between two key poses whose weights are respectively w_i and w'_j. This distance takes into account the distance of the features and the relevance of the specific match of key poses. In short, priority is given to matches between very discriminative key poses: if their distance is low, it should be decreased; if it is high, it should be increased. The distance is obtained as follows:

1. Let *average_distance* be the average Euclidean distance between features obtained at the training stage. The signed deviation of the distance between the key pose features is defined as:

$$dev(i, j) = |kp_i - kp'_j| - average\_distance \quad . \tag{8}$$

2. The relevance $rel(i, j)$ of the match of key poses is obtained by correlating the deviation of the feature distance with the weights of the key poses. In this way, feature distances close to average are considered as irrelevant:

$$rel(i, j) = |dev(i, j) * w_i * w'_j| \quad . \tag{9}$$

3. This relevance is added to the feature distance in order to obtain the key pose distance:

$$d(kp_i, kp'_j) = |kp_i - kp'_j| + v \;\; rel(i, j) \quad , \tag{10}$$

where the value of v is decided based upon the desired behavior, which is summarized in the following table.

Table 1. Value of v based on the pairing of key poses and the signed deviation. Key poses are understood as ambiguous when their weights are under 10%, and as discriminative when they are over 90%. (These values have been chosen empirically.)

Signed deviation	Pairing	v
$dev(i,j) < 0$	*both discriminative*	-1
$dev(i,j) > 0$	*both discriminative*	$+1$
any	*both ambiguous*	-1
any	*a discriminative and an ambiguous*	$+1$

In discriminative pairings, i.e. key poses with high weights, matches are very significant, as these are the key poses that uniquely distinguish one action from another. Therefore, their relevance will decrease the key pose distance if the signed deviation is below zero, or increase it if it is above zero. On the other hand, when both key poses are ambiguous, i.e. they have low weights, these pairings should not be considered as important as the rest. That is why a negative relevance is used so as to reduce the influence of their distance. Lastly, a pairing of a discriminative key pose and an ambiguous one will be unfavored, since both should be able to find matches with similar weights. In the remaining cases of pairings which are not shown in Table 1, the same value as for *both discriminative* pairings is used.

In conclusion, this *weight tuning scheme* allows us to influence the behavior of the sequence matching algorithm by favoring those matches from which we know that they are important, and disfavoring those which are not.

Finally, evaluating the DTW distance between the test sequence and the previously learned training sequences, the nearest neighbor sequence of key poses is found for each view, and the label of the closest match supplies the final result. Hence, the best matching view is used and single-view action recognition is supported if necessary.

6 Experimentation

In order to analyse the performance of the proposed multi-view action recognition method, the MuHAVi [21] dataset has been chosen. This dataset includes multi-view images of a resolution of 720x576 px in complex background and lighting conditions. Specifically, the MuHAVi-MAS (Manually Annotated Silhouettes) is used as it provides silhouettes for two different views (front-side and 45° angle). Two actors were recorded performing 14, or 8 actions in its simplified version. The authors of this dataset introduced two additional tests so as to verify the actor- and view-invariance of the method. These tests give further insight of the method's behavior when training and testing conditions differ. Only the actor-invariance test has been applied because the introduced view-invariance test does not support multi-view learning.

Furthermore, not only the achieved results are compared with those of similar state-of-the-art methods, but also a different type of data fusion from multiple

views has been developed with the purpose of verifying the chosen *Model Fusion* approach. In this sense, *Feature Fusion* based on concatenation of features as proposed in [13] is used to obtain a multi-view pose representation. Simplifying the proposed method to only one view, made up of multi-view pose representations, the algorithm is used as before. Lastly, results without any type of data fusion are obtained. In this case, the system is fed with video sequences ignoring from which view they come from, and at testing each view is considered as an independent test sequence.

The constant parameters of the presented method L, the feature size, and K, the number of key poses per action class (and per view in the case of *Model Fusion*), are detailed for the best results achieved.

6.1 Leave-one-sequence-out Cross Validation

In *leave-one-sequence-out* cross validation, one fold per sequence is executed, i.e. all but one sequence are used for training and the remaining one is used for testing. This procedure is repeated for each possible test sequence, and the accuracy scores are averaged. Figure 2 shows the obtained confusion matrix for MuHAVi-14 in which it can be seen that only 4 sequences are misclassified, achieving a final recognition rate of 94.1%. Furthermore, the matrix shows that errors are mostly made between very similar actions as *StandupLeft* and *StandupRight*. On MuHAVi-8 (see figure 3) only one sequence is misclassified and a final recognition rate of 98.5% is reached.

	CollapseLeft	CollapseRight	GuardToKick	GuardToPunch	KickRight	PunchRight	RunLeftToRight	RunRightToLeft	StandupLeft	StandupRight	TurnBackLeft	TurnBackRight	WalkLeftToRight	WalkRightToLeft
CollapseLeft	4/4													
CollapseRight		4/4												
GuardToKick			8/8											
GuardToPunch			1/8	7/8										
KickRight					8/8									
PunchRight						8/8								
RunLeftToRight							3/4	1/4						
RunRightToLeft								4/4						
StandupLeft									1/2	1/2				
StandupRight										4/4				
TurnBackLeft											2/2			
TurnBackRight												3/4	1/4	
WalkLeftToRight													4/4	
WalkRightToLeft														4/4

Fig. 2. Confusion matrix of the *leave-one-sequence-out* cross validation on the MuHAVi-14 dataset

Table 2 shows the results that have been achieved with the *Feature Fusion* approach or without considering any multi-view recognition. It can be seen that

	Collapse	Guard	KickRight	PunchRight	Run	Standup	TurnBack	Walk
Collapse	8/8							
Guard		16/16						
KickRight			8/8					
PunchRight				8/8				
Run					8/8			
Standup						6/6		
TurnBack							5/6	1/6
Walk								8/8

Fig. 3. Confusion matrix of the *leave-one-sequence-out* cross validation on the MuHAVi-8 dataset

Model Fusion clearly outperforms these methods, and that a significant improvement is obtained when multiple views are learned explicitly at the model level. Moreover, in comparison with the baseline and the latest and highest recognition rates achieved by other state-of-the-art methods, our approach presents, to the best of our knowledge, the best result so far on the MuHAVi-14 dataset.

Table 2. Comparison of our results with similar state-of-the-art approaches on the MuHAVi dataset (all use *leave-one-sequence-out* cross validation)

Approach	**MuHAVi-14**		**MuHAVi-8**	
Singh et al. [21] (baseline)		82.4%		97.8%
Cheema et al. [22]		86.0%		95.6%
Martínez-Contreras et al. [23]		-		98.4%
Eweiwi et al. [24]		91.9%		98.5%
Without Fusion	$(L = 600, K = 120)$	85.3%	$(L = 350, K = 60)$	95.6%
Feature Fusion	$(L = 200, K = 140)$	92.6%	$(L = 300, K = 100)$	97.1%
Model Fusion	$(L = 450, K = 60)$	94.1%	$(L = 250, K = 75)$	98.5%

6.2 Novel Actor Test

In this section, actor-invariance is tested. For this purpose, all the sequences of one actor are used for training, whereas the sequences of the second actor are used for testing. This test is executed twice, interchanging the training and the testing groups and averaging the accuracy scores. Table 3 shows the results of the *Novel Actor* test. Again, *Model Fusion* achieves significantly better results than the other methods. Interestingly, *Feature Fusion* performs worse than the single-view recognition. This can be attributed to the increased actor-variance that results of using the multi-view pose representation. These results outperform the currently available recognition rates by 8.9% and 10.3% respectively.

The presented method performs specially well in this test, because of the performed shift from sequences of pose representations to sequences of key poses.

Table 3. Comparison of results of the MuHAVi *Novel Actor* test

Approach	**MuHAVi-14**		**MuHAVi-8**	
Singh et al. [21] (baseline)		61.8%		76.4%
Cheema et al. [22]		73.5%		83.1%
Eweiwi et al. [24]		77.9%		85.3%
Without Fusion	$(L = 200, K = 80)$	81.6%	$(L = 300, K = 60)$	92.6%
Feature Fusion	$(L = 200, K = 100)$	80.9%	$(L = 200, K = 100)$	91.2%
Model Fusion	$(L = 450, K = 60)$	86.8%	$(L = 250, K = 75)$	95.6%

This moves the test data domain to our domain of key poses and constitutes an essential step in the process, as noise and possible dissimilarities between actors are filtered.

6.3 Temporal Evaluation

As stated beforehand, our work has been driven by the ambition of creating a multi-view action recognition method which could deal with both the increased complexity of multi-view learning, and the necessity of an adequate recognition speed in order to perform real-time action recognition. This guided the decisions that have been taken about the design of the multi-view action recognition method whose temporal performance is tested in this section.

The temporal evaluation of the presented method has been performed on a standard PC with an Intel Core 2 Duo CPU at 3 GHz with Windows 7 64-bit. The method has been implemented using the OpenCV library [25] and the .NET Framework. All the necessary processing stages have been measured taking the binary silhouette as input and going trough contour extraction, feature extraction and multi-view learning by *Model Fusion* or classification by means of sequence matching. No specific hardware optimizations have been performed.

Running the MuHAVi-14 benchmark, each sequence is processed in 1.14 s achieving a recognition speed of 51 FPS. As MuHAVi-8 presents fewer classes, the recognition speed rises to 66 FPS, i.e. 0.88 s per sequence.

It is worth mentioning that the presented approach also proofs to be efficient at the training stage. An average training speed of 39 FPS and 50 FPS has been measured on MuHAVi-14 and MuHAVi-8 respectively.

7 Conclusion and Discussion

In this paper, a multi-view action recognition method based on a bag-of-key-poses has been presented. A simple contour-based feature allows us to obtain a very efficient pose representation whose most characteristic instances for each action class and view are learned by means of the bag-of-key-poses. Sequence matching finally leads to consider the temporal evolution between key poses. The method obtains highly stable and accurate results and establishes new reference recognition rates for the MuHAVi dataset. Furthermore, real-time suitability is

shown in the temporal evaluation because the method performs above video frequency.

Future work includes experimentation on other datasets with more camera views. This will foreseeably affect the method's performance and may require adjustments in the learning scheme. Nevertheless, the ideal number of camera views should be determined for each application scenario since FOV conditions change between indoor and outdoor spaces.

Acknowledgements. This work has been partially supported by the Spanish Ministry of Science and Innovation under project "Sistema de visión para la monitorización de la actividad de la vida diaria en el hogar" (TIN2010-20510-C04-02) and by the European Commission under project "caring4U - A study on people activity in private spaces: towards a multisensor network that meets privacy requirements" (PIEF-GA-2010-274649). Alexandros Andre Chaaraoui acknowledges financial support by the Conselleria d'Educació, Formació i Ocupació of the Generalitat Valenciana (fellowship ACIF/2011/160).

References

1. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
2. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vis. Image Underst. 115(2), 224–241 (2011)
3. Holte, M.B., Tran, C., Trivedi, M.M., Moeslund, T.B.: Human action recognition using multiple views: a comparative perspective on recent developments. In: Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, J-HGBU 2011, pp. 47–52. ACM, New York (2011)
4. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3), 257–267 (2001)
5. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Comput. Vis. Image Underst. 104(2), 249–257 (2006)
6. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, pp. 1395–1402 (2005)
7. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64, 107–123 (2005)
8. Oikonomopoulos, A., Patras, I., Pantic, M.: Spatiotemporal salient points for visual recognition of human actions. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 36(3), 710–719 (2005)
9. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional SIFT descriptor and its application to action recognition. In: Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA 2007, pp. 357–360. ACM, New York (2007)
10. İkizler, N., Duygulu, P.: Human Action Recognition Using Distribution of Oriented Rectangular Patches. In: Elgammal, A., Rosenhahn, B., Klette, R. (eds.) Human Motion 2007. LNCS, vol. 4814, pp. 271–284. Springer, Heidelberg (2007), http://dx.doi.org/10.1007/978-3-540-75703-0_19

11. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. 43(3), 16:1–16:43 (2011)
12. Määttä, T., Härmä, A., Aghajan, H.: On efficient use of multi-view data for activity recognition. In: Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2010, pp. 158–165. ACM, New York (2010)
13. Wu, C., Khalili, A.H., Aghajan, H.: Multiview activity recognition in smart homes with spatio-temporal features. In: Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2010, pp. 142–149. ACM, New York (2010)
14. Naiel, M.A., Abdelwahab, M.M., El-Saban, M.: Multi-view human action recognition system employing 2DPCA. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 270–275 (2011)
15. Cilla, R., Patricio, M.A., Berlanga, A., Molina, J.M.: A probabilistic, discriminative and distributed system for the recognition of human actions from multiple views. Neurocomputing 75(1), 78–87 (2012); Brazilian Symposium on Neural Networks (SBRN 2010), International Conference on Hybrid Artificial Intelligence Systems (HAIS 2010)
16. Yan, S.M.P., Khan, Shah, M.: Learning 4D action feature models for arbitrary view action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–7 (2008)
17. Canton-Ferrer, C., Casas, J.R., Pardas, M.: Human model and motion based 3D action recognition in multiple view scenarios. In: Conf. on 14th European Signal Processing, Italy, pp. 1–5 (2006)
18. Suzuki, S., Abe, K.: Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing 30(1), 32–46 (1985)
19. Ángeles Mendoza, M., Pérez de la Blanca, N.: HMM-Based Action Recognition Using Contour Histograms. In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (eds.) IbPRIA 2007. LNCS, vol. 4477, pp. 394–401. Springer, Heidelberg (2007), http://dx.doi.org/10.1007/978-3-540-72847-4_51
20. Dedeoğlu, Y., Töreyin, B., Güdükbay, U., Çetin, A.: Silhouette-Based Method for Object Classification and Human Action Recognition in Video. In: Huang, T.S., Sebe, N., Lew, M., Pavlović, V., Kölsch, M., Galata, A., Kisačanin, B. (eds.) HCI/ECCV 2006. LNCS, vol. 3979, pp. 64–77. Springer, Heidelberg (2006), http://dx.doi.org/10.1007/11754336_7
21. Singh, S., Velastin, S.A., Ragheb, H.: Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 48–55 (2010)
22. Cheema, S., Eweiwi, A., Thurau, C., Bauckhage, C.: Action recognition by learning discriminative key poses. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1302–1309 (2011)
23. Martínez-Contreras, F., Orrite-Urunuela, C., Herrero-Jaraba, E., Ragheb, H., Velastin, S.A.: Recognizing human actions using silhouette-based HMM. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009, pp. 43–48 (2009)
24. Eweiwi, A., Cheema, S., Thurau, C., Bauckhage, C.: Temporal key poses for human action recognition. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1310–1317 (2011)
25. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)

Bayesian Fusion of Ceiling Mounted Camera and Laser Range Finder on a Mobile Robot for People Detection and Localization

Ninghang Hu[1], Gwenn Englebienne[1], and Ben J.A. Kröse[1,2]

[1] University of Amsterdam
[2] Amsterdam University of Applied Science

Abstract. Robust people detection and localization is a prerequisite for many applications where service robots interact with humans. Future robots will not be stand-alone any more but will operate in smart environments that are equipped with sensor systems for context awareness and activity recognition. This paper describes a probabilistic framework for the fusion of data from a laser range finder on a mobile robot and an overhead camera fixed in a domestic environment. The contribution of the framework is that it enables seamless integration with other sensors. For tracking multiple people it is possible to use a probabilistic particle filter tracker. We show that the fusion improves the results of the individual subsystems.

1 Introduction

As the baby boom generation is coming to retirement age, the number of elderly citizens over 60 years of age is expected to grow further to a proportion of 1 out of 3 by the year 2030. Alongside this growth in the elderly population, we face short and long-term labor shortages, especially in the health-care sector. Robots may offer a solution for making elderly care affordable by using them for physical [9], cognitive [12] or social [16] support. All these studies share a common foundation that the robots interact intensively with humans, and locations, of both the person and the robot, are estimated robustly.

Sensing systems for robot localization or people localization are usually mounted either on the robot or are fixed in the environment. In this paper we describe a probabilistic framework for the *fusion* of data from robot and fixed sensors. Here we restrict ourselves to a laser scanner on the robot and an overhead camera fixed in the room. The contribution of our work is that by mapping all information into a probabilistic model, the system can be easily extended with other sensors such as multiple cameras or RGB-D cameras, and is robust to the absence of sensors.

In the next sections we will describe related work and systems. In section 4 our own model will be introduced. The sections after that will describe the likelihood functions for the laser on the robot and the camera in the room. In section 7 we describe the experiments and results. We conclude with a discussion on the method and results.

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 41–51, 2012.

2 Related Work

There is a long tradition of research in the field of people detection and localization in robot applications. Many studies concentrate on people detection using the sensors on the mobile robot. Relatively simple sensors such as laser range finders were used for detection and localization [11,15]. People are extracted from range data as single blobs or found by merging nearby point clusters that correspond to legs. Probabilistic techniques such as multi-hypothesis trackers are used for tracking multiple objects [1].

Instead of using the laser range systems on the robots, vision systems have also been used for people detection. Since robot-mounted cameras are moving, the detection cannot be based on background modeling methods, and local characteristics such as color histograms or local features have been used [14,17]. To make detection more robust, the fusion of different modalities of robot sensors is suggested. Leg detection by laser range finders in combination with face detection has shown to be more robust than individual modalities [10,2]. In [18], Viola-Jones type of visual detectors are used to recognize body parts and are combined with laser range data.

However, future robots will operate in smart homes that are equipped with sensors, and it seems obvious to use these sensors also for person detection. One advantage is that the system may be more robust: noise or deviations in a sensor may be detected and corrected. Another advantage is that the robot does not need to keep monitoring the persons all the time. The robot may be required to finish other tasks from time to time, rather than allocating its resources to the task of tracking each person all the time.

Person tracking systems that are mounted in domestic environments are usually based on vision systems, although there are some exceptions using laser range finders [8] or speech [6]. Overhead cameras are often used, which are usually mounted very high, and have a very wide angle of view, covering most of the areas in the room. Since they look down from above, it turns out that human users are less likely to be occluded compared with cameras mounted on the side. An application in a kids playroom is given in [3].

In our set-up we combine an overhead camera with the laser range finder on the robot. In order to have a sound probabilistic framework we build on the approach of [7], who uses a probabilistic foreground segmentation with a template based detection. The result is a posterior distribution on the locations of the persons in the room. This is combined with a distribution based on the laser range finder.

3 System Overview

Our proposed system is used to detect and localize the elderly people in chores of robot home-care. With our system, the robot is able to obtain accurate locations of the users in the room, and thereby it can interact with the human users precisely. The robot we use possesses multiple on-board sensors, including a Kinect camera, a stereo camera, and a laser range finder.

In the recent work, most of the robots are designed for following the targets. These approaches, therefore, require that the users are always in the range of the robot sensors. In the case of home care, however, the robot moves around in the room to execute

a variety of tasks, and at some points the robot sensors will loss the track of the human user, e.g. the robot is asked to get an object that is in an opposite direction to the user. To overcome such a problem and enable continuous human localization, we adopt an ambient camera and mount it on the ceiling of the room. The advantage of the ambient camera is twofold: (1) that it gives a top view of the whole room, and (2) that people in the room are less likely to be occluded compared with the robot cameras. Since it covers the whole area of the room, the ceiling-mounted camera is able to localize persons continuously when the users are present in the room, so that when the robot sensors fail to detect the users, the ambient camera is still able to report the correct location to the system. Besides, the robot sensors and the ambient camera observe the persons from different directions, giving complementary cues for the human detection and localization. The fusion system can, therefore, obtain a better estimate of the location of the users compared with the approaches using single modality.

To combine the robot sensors and the ambient camera, we propose a Bayesian fusion framework. Next, we formulate the problem and introduce our fusion framework.

4 Probabilistic Fusion Framework

The Bayesian approach provides an elegant way of fusing between different sensor sources as well as dealing with noise and uncertainty in sensor measurements [13].

Assume I_R is the observed data from the robot sensor, and I_C is observed from the ambient camera, *i.e.* the overhead camera. Given I_R and I_C, we aim to find a robust estimation of the location of multiple persons L_H, the location of the robot L_R, and the orientation of the robot θ_R. In the context of a Bayesian framework, the posterior distribution $P(L_R, L_H, \theta_R | I_R, I_C)$ is the target we would like to know by the end.

Using the Bayesian Theorem, the posterior probability can be derived as

$$P(L_R, L_H, \theta_R | I_R, I_C) \propto P(I_R, I_C | L_R, L_H, \theta_R) P(L_R, L_H, \theta_R) \tag{1}$$

where $P(L_R, L_H, \theta_R) = p(L_R)p(L_H)p(\theta_R)$ is the prior distribution that is known before the sensory data is observed. These priors can be estimated either from separate training data, or from prior knowledge of the problem. In our case, we simply assume a uniform distribution over the ground area of the floor, and a uniform distribution over the angles of the orientation. $P(I_R, I_C | L_R, L_H, \theta_R)$ is the likelihood.

By assuming I_R and I_C are measured independent with separate sensors, and I_C is not dependent on the rotation of the robot θ_R, the likelihood probability of Equ. 1 can be decomposed as

$$P(I_R, I_C | L_R, L_H, \theta_R) = P(I_R | L_R, L_H, \theta_R) P(I_C | L_R, L_H) \tag{2}$$

where $P(I_R | L_R, L_H, \theta_R)$ is the likelihood of generating the image I_R given the combination of L_R, L_H, and θ_R, while $P(I_C | L_R, L_H)$ represents the likelihood of the ambient camera that generates the observation I_C.

Again, our goal is to find the optimal combination of L_R^*, L_H^* and θ_R^* that maximizes the posterior distribution $P(L_R, L_H, \theta_R | I_R, I_C)$, which is a typical MAP problem that can be solved by particle filtering [5].

The camera likelihood $P(I_C|L_R,L_H)$ is used as the proposal distribution to sample particles, and the particles are weighted by the corresponding likelihood of the laser data $P(I_R|L_R,L_H,\theta_R)$. The optimal combination L_R^*, L_H^* and θ_R^* is considered as the particle that holds the highest weight. In a Bayesian framework, however, we find the expectation of the parameter values rather than the most probable value. Therefore, rather than choosing one particle that maximizes the joint distribution, we compute the solution as a weighted sum of all the particles.

The remaining is to compute the two likelihood terms in Equ. 2. In the following two sections, we introduce the methods of estimating the two likelihood items separately. Here, we will focus on modeling the likelihood of the robot sensor. For the camera, we adopt the algorithm from [7].

5 Measuring Likelihood of Robot Sensors

In our data fusion framework, the state is to be estimated is a triplet of $\{L_R,L_H,\theta_R\}$. The likelihood of the robot sensors measures the probability of generating the observation I_C rather than all the observations that can be possibly generated from such a triplet, given such a state triplet, *i.e.* $P(I_R|L_R,L_H,\theta_R)$.

In this paper, we adopt the Laser Range Finder as our robot sensor. The Laser Range Finder scans in a plain and detects the distance to the objects in range. In the context of human localization in a home setting, the detected objects can either be objects that exist in the room or be part of the human in the room. In this paper, we use the background model and the human model, respectively, to model the probability that a region is occupied by either of these two objects. Then we can compute the occupancy map of the room, *i.e.* the probability that the area is occupied by either the background object or by a human.

The occupancy maps are used to estimate the probability of the robot sensor generating a certain set of observations, *i.e.* the robot sensor's likelihood.

5.1 Probabilistic Background Model

To find out what the room looks like in terms of background obstacles, the robot is first driven around to build a background model of the room.

For each time stamp, the robot sensor fires a set of laser beams $l = \{l_1,l_2,l_3,...\}$. Whenever there is an object in the way, the laser is reflected back to the base and thereby the distance to the background objects is detected. Given the coarse location of the robot, we are able to find the approximate locations of these laser detections. But due to the uncertainty in the location of the robot as well as the noise in laser data, these locations are not fully reliable. Therefore, simply giving a Boolean answer to the occupation of the local region is not an elegant solution, and a probabilistic way of modeling the background is required.

In our approach, the ground plane is first discretized into small cells of equal size. We denote k as the index of the cell on the ground plane. Then for each cell k, we aim to estimate the probability that the cell is occupied by a part of the background. Collectively, these probabilities form the background model $P_b(k)$.

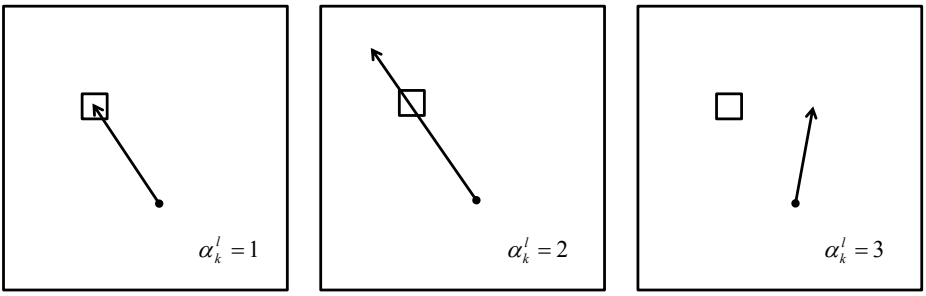


Fig. 1. The relation between a laser beam and a cell can be summarized into three patterns. In the left pattern $\alpha_k^l = 1$, the laser is blocked by the cell, referring that the cell is occupied by certain background objects. The middle pattern $\alpha_k^l = 2$ shows the laser has passed through the cell, indicating the cell is empty. As for the third pattern, however, the laser beam is blocked before it reaches the cell. Therefore, no clue about whether the cell is occupied can be deduced from the third pattern.

In this paper, the background model $P_b(k)$ is measured as the number of times the laser scanner observes an occupied cell normalized by the number of times that the cell is in the range of the laser scanner. To formalize the problem, we define three patterns that can be observed given a scan l and a cell k, see Fig. 1. We use a random variable α_k^l to denote the index of the three patterns. The first pattern refers that the cell k is detected by l as an occupied cell. The second pattern denotes that the cell is observed as an empty cell. As for the third pattern, no information about the cell can be inferred since the cell is either occluded by other background objects that is in front of the cell, or the laser is not fired in the direction of the cell. Therefore, the third pattern does not contribute to the background model while only the first two do. Next, we estimate the background model by

$$P_b(k) = \frac{\sum_l \delta(\alpha_k^l - 1)}{\sum_l \delta(\alpha_k^l - 1) + \delta(\alpha_k^l - 2)} \tag{3}$$

where δ is a Kronecker delta function, and the equation sums over all the lasers that pass through the cell k.

5.2 Learning Human Model

The human model P_h reflects how the human looks like from the robot sensors in the world frame. It is learned by accumulating the laser points that locate in a small region around the center of the person. Each pixel in such a region holds a value indicating the probability that the cell is occupied by the person, *i.e.* a higher value means the cell is more likely to be detected by the robot sensor due to the occurrence of the human.

Similar to training the background model, we learn the human model P_h by calculating the number of laser beams that either have a positive detection at the cell or pass through the cell. Again, we adopt the Equ. 3 for computing the human model P_h.

Given the person locating in cell k, the local human model P_h can be translated into the world frame to generate a human model map $P_h(k)$.

5.3 Occupancy Map

Knowing the background model and the human model, we are able to compute the probability of occupancy for each of the cells on the ground plane. Note that the cell cannot be occupied by both the human and the background obstacle at the same time, therefore the occupancy map is computed as

$$\omega_k = \frac{P_b(k)\tilde{P}_h(k) + P_h(k)\tilde{P}_b(k)}{1 - P_b(k)P_h(k)} \tag{4}$$

where

$$\tilde{P}(k) = 1 - P(k) \tag{5}$$

5.4 Likelihood of Laser Range Finder

The likelihood of the Laser Range Finder denotes the probability of generating the current observation given the state $\{L_R, L_H, \theta_R\}$. I_R represents a vector of the laser range data. Assume I_R contains N independent measurements $\{i_R^1, i_R^2, ..., i_R^n, ..., i_R^N\}$. Suppose the direction of the range measurement i_R^n is defined by θ_R^n. Therefore

$$P(I_R|L_R, L_H, \theta_R) = \prod_{n=1}^{N} P(i_R^n|L_R, L_H, \theta_R^n) \tag{6}$$

L_R and θ_R^n define a robot at the location L_R, and the robot fires a laser beam in the direction of θ_R^n. L_H refers to the location of multiple persons.

Suppose the laser beam i_R^n passes through a set of cells in a straight line, e.g. $\{c_1, c_2, ..., c_{m-1}\}$, and then it detects a certain object at the cell c_m. c_M denotes the maximal range that the laser can reach. See Fig. 2. Then the probability of obtaining a detection at cell c_m rather than the other locations can be computed by

$$P(i_R^n|L_R, L_H, \theta_R^n) = \frac{\omega_{c_m} \prod_{i=1}^{m-1} \tilde{\omega}_{c_i}}{\sum_{j=1}^{M} \omega_{c_j} \prod_{i=1}^{j-1} \tilde{\omega}_{c_i}} \tag{7}$$

Since multiplications of the probabilities can result in very small numbers which lead to floating point overflows, we compute the log-likelihood instead

$$\mathscr{L}(i_R^n|L_R, L_H, \theta_R^n) = \lambda_m - \sum_{j=1}^{M} \lambda_j \tag{8}$$

where

$$\lambda_m = log(\omega_{c_m}) + \sum_{i=1}^{m-1} log(\tilde{\omega}_{c_i}) \tag{9}$$

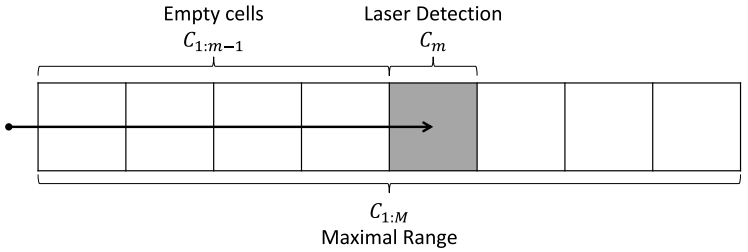


Fig. 2. The laser beam (Arrow) passes through $m-1$ empty cells and finally reaches the cell at C_m. The maximal range of the laser covers M cells.

6 Likelihood of Ceiling Mounted Camera

The likelihood of over head camera is computed the same way as in [7]. Assuming the pixels are independent from each other given the image taken by the ceiling mounted camera, the likelihood $P(I_C|L_R,L_H)$ can be derived as

$$P(I_C|L_R,L_H) = \prod_{n=1:N} P(i_C^n|L_R,L_H) \tag{10}$$

We build a specific polyhedron to model the 3D shape of both the human and the robot. Given the location of the human L_H and the robot L_R, the polyhedrons are projected into the image space, generating a foreground mask $\mathscr{M}$. For each pixel location $P(i_C^n)$ on the image, we look up in the mask and use $\mathscr{M}_n$ to determine whether the pixel is a part of the foreground or background. Then the likelihood can be computed as

$$P(i_C^n|L_R,L_H) = P_f(i_C^n)\mathscr{M}_n + P_b(i_C^n)(1-\mathscr{M}_n) \tag{11}$$

where $P_b(i_C^n)$ is the background model which is learned beforehand using the background images. $P_f(i_C^n)$ is the foreground model, and in our case we apply a uniform distribution over the colors.

7 Experiment and Results

The proposed data fusion framework was evaluated on data collected with a Nomad robot platform and an overhead camera, see Fig. 3. The overhead camera is mounted centrally on the ceiling and gives a panoramic view of the room. The frames that are captured with the camera are highly distorted due to the fish-eye effect. The camera's lens parameters are calibrated with the OpenCV module [4].

On the Nomad robot platform, a Laser Range Finder, a Kinect camera and a stereo camera are mounted on the robot. For the present experiments, we restrict ourselves to test the framework by using the Laser Range Finder, mounted at a height of 20 cm. The robot is remote-controlled and manually driven around in the room. The robot records its odometry information by measuring the rotations of its two wheels. The odometry data are then adopted for generating the orientation and location of the robot.

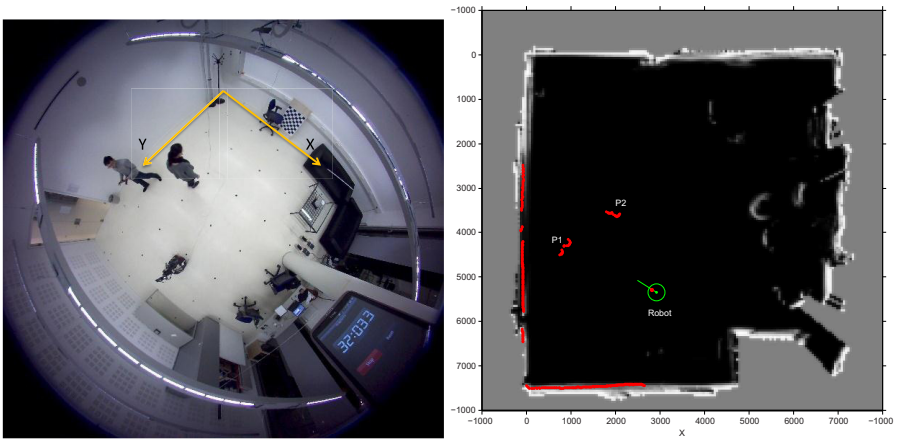


Fig. 3. An overview of the experiment room and the observed data. Left: captured by the over head camera; Right: laser detection points (red dots).

The Nomad robot runs on the Robot Operating System (ROS), and all data captured on the robot site is time stamped in ROS.

The exact time stamp of each frame collected with the overhead camera is obtained by means of a stopwatch mounted close to the camera. We use a nearest-neighbors classifier to recognize digits in the image to recover the time stamp. We synchronized the robot sensors and the overhead camera based on specific time points, where an event (e.g. the puncturing a balloon in front of the Laser Range Finder) was observed by both the robot sensor and the overhead camera.

The ground plane is subdivided into small cells of 50×50mm. In a first training run, the robot was remote-controlled to generate the background model. Second, the human model is trained according to Equ. 3. During testing, the two models are combined probabilistically into an occupancy map given the particles, as depicted in Fig. 4. Here each pixel of the occupancy map reflects the probability that that location is occupied, either by a person or by a background object in the room.

We evaluate the systems by measuring the Euclidean distance between the detection results and the ground truth locations of persons. In this paper, three localization approaches are tested and compared: a) with a single Laser Range Finder; b) with a single over head camera; c) with our proposed fusion framework. We evaluate the proposed system and the single modality approaches on 165 camera frames together with synchronized laser data. For each of these frames, volunteers manually annotated the locations of the persons in the ground plane, based on physical markers that were positioned on the floor during the recording, and these markers were used as reference to compute the ground truth location.

A particle sampling approach is applied both in the single laser and the data fusion approach. An equal number of 800 particles are sampled. Due to the fact that humans are not likely to be too close to each other, we define the safe distance between two persons as 500mm. We incorporate such assumption to reduce the space when sampling

Fig. 4. The occupancy map is generated from combining the human model and the background model. For each set of locations of persons, *i.e.* a particle, an occupancy map is estimated. Left: Human model. Middle: Background model. Right: Occupancy Map given the hypothetical location of persons (green crosses).

particles, *i.e.* the sampled point is always at least 500mm away from each of the points the previous sample set.

The single laser approach detects the foreground laser points by set a threshold to their probability in the background model. The threshold in our experiment is empirically set to 0.3. The particles are sampled from the foreground laser points with a Normal distribution on the location of the points. The weights are assigned by the likelihood of the laser data given the particles, and they are quantized in the sub-divided cells on the ground plane according to the locations of the particles. The human is then localized by recursively finding the cell that has the largest sum of weights as in [7].

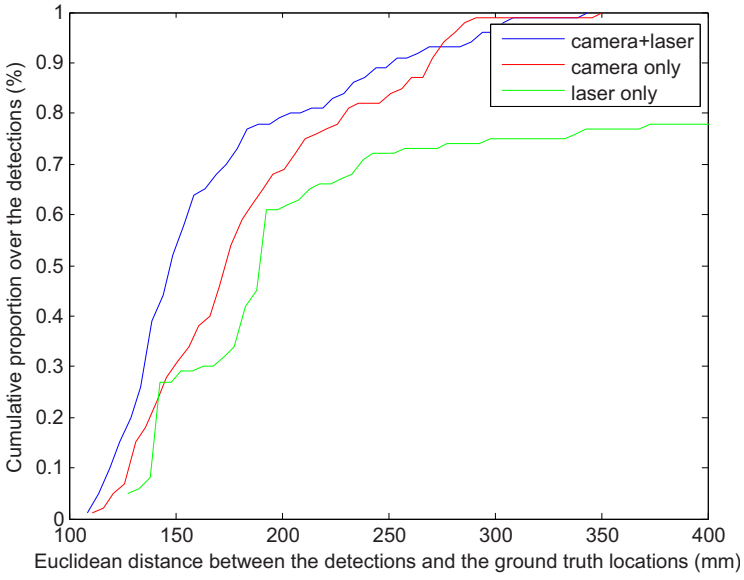


Fig. 5. Comparing the proposed data fusion approach and the single modality approach

In the approach with a single camera, we adopt the human detection algorithm from [7]. For each candidate location of the persons on the ground plane, the likelihood of the camera frame is measured. The locations of the multiple persons are found by choosing the locations that maximize the likelihood of the camera image.

The proposed approach combines the over head camera and the robot Laser Range Finder in a probabilistic Bayesian framework. After persons are localized with the single camera, the particles are sampled around the location of the persons with a Normal distribution. These particles are then weighted by the likelihood of the laser observations. The final detection is computed by the weighted sum of the particles that are sampled from the same person.

Fig. 5 shows the detection results of our data fusion system comparing with the approach using single modality. The proposed fusion system consistently outperforms the single-camera and the single-laser approach, and approximately 80 percent of the detections are less than 200 mm from the ground truth location. In contrast, only 70% of the camera-only detections and 27% of the laser-only detections are within such distance of the ground truth.

8 Conclusion and Future Work

We have proposed a novel probabilistic fusion framework for the localization of humans using ambient cameras and robot-mounted Laser Range Finders. Our experiments show substantial improvements in the accuracy of the localization, thus enabling more precise interaction between robot and humans. Due to its probabilistic nature, our framework can deal with occlusions and the absence of measurements in a principled way. As a result, the localization of humans is more robust, and natural interaction becomes possible even in challenging conditions.

In our current experimental work, the orientation and the location are not considered as part of the particle, but only the location of multiple persons are sampled. But we expect the performance can be improved by incorporating robot location and orientation into particles. We plan to specifically address occlusions and missed detections in one of the sensors. We will also extend the method to use more and different sensors, including the robot-mounted Kinect camera, as well as multiple overhead cameras.

Acknowledgments. This research is funded by the EU FP7-287624 Acceptable robotiCs COMPanions for AgeiNg Years (ACCOMPANY) and by the SIA project BALANCE-IT.

References

1. Arras, K.O., Grzonka, S., Luber, M., Burgard, W.: Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In: IEEE International Conference on Robotics and Automation, ICRA 2008, pp. 1710–1715. IEEE (2008)
2. Bellotto, N., Hu, H.: Vision and laser data fusion for tracking people with a mobile robot. In: IEEE International Conference on Robotics and Biomimetics, ROBIO 2006, pp. 7–12. IEEE (2006)

3. Bobick, A.F., Intille, S.S., Davis, J.W., Baird, F., Pinhanez, C.S., Campbell, L.W., Ivanov, Y.A., Schütte, A., Wilson, A.: The kidsroom: A perceptually-based interactive and immersive story environment. Presence 8(4), 369–393 (1999)
4. Bradski, G., Kaehler, A.: Learning OpenCV: Computer vision with the OpenCV library. O'Reilly Media (2008)
5. Cai, Y., de Freitas, N., Little, J.J.: Robust Visual Tracking for Multiple Targets. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 107–118. Springer, Heidelberg (2006)
6. Checka, N., Wilson, K.W., Siracusa, M.R., Darrell, T.: Multiple person and speaker activity tracking with a particle filter. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, vol. 5, pp. V–881. IEEE (2004)
7. Englebienne, G., Kröse, B.J.A.: Fast bayesian people detection. In: Proceedings of the 22nd Benelux AI Conference, BNAIC 2010 (2010)
8. Fod, A., Howard, A., Mataric, M.A.J.: A laser-based people tracker. In: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA 2002, vol. 3, pp. 3024–3029. IEEE (2002)
9. Graf, B.: Reactive navigation of an intelligent robotic walking aid. In: Proceedings of the10th IEEE International Workshop on Robot and Human Interactive Communication 2001, pp. 353–358. IEEE (2001)
10. Kleinehagenbrock, M., Lang, S., Fritsch, J., Lomker, F., Fink, G.A., Sagerer, G.: Person tracking with a mobile robot based on multi-modal anchoring. In: Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication 2002, pp. 423–429. IEEE (2002)
11. Kluge, B., Kohler, C., Prassler, E.: Fast and robust tracking of multiple moving objects with a laser range finder. In: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA 2001, vol. 2, pp. 1683–1688. IEEE (2001)
12. Pineau, J., Montemerlo, M., Pollack, M., Roy, N., Thrun, S.: Towards robotic assistants in nursing homes: Challenges and results. Robotics and Autonomous Systems 42(3), 271–281 (2003)
13. Punska, O.: Bayesian approaches to multi-sensor data fusion. Cambridge University, Cambridge (1999)
14. Schlegel, C., Illmann, J., Jaberg, H., Schuster, M., Wörz, R.: Vision based person tracking with a mobile robot. In: British Machine Vision Conference, pp. 418–427 (1998)
15. Song, X., Cui, J., Wang, X., Zhao, H., Zha, H.: Tracking interacting targets with laser scanner via on-line supervised learning. In: IEEE International Conference on Robotics and Automation, ICRA 2008., pp. 2271–2276. IEEE (2008)
16. Wada, K., Shibata, T.: Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house. IEEE Transactions on Robotics 23(5), 972–980 (2007)
17. Zajdel, W., Zivkovic, Z., Krose, B.J.A.: Keeping track of humans: Have i seen this person before? In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, ICRA 2005, pp. 2081–2086. IEEE (2005)
18. Zivkovic, Z., Krose, B.: Part based people detection using 2d range data and images. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2007, pp. 214–219. IEEE (2007)

Using Speech Data to Recognize Emotion in Human Gait

Angelica Lim and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto, Japan 606-8501
{angelica,okuno}@kuis.kyoto-u.ac.jp

Abstract. Robots that can recognize emotions can improve humans' mental health by providing empathy and social communication. Emotion recognition by robots is challenging because unlike in human-computer environments, facial information is not always available. Instead, our method proposes using speech and gait analysis to recognize human emotion. Previous research suggests that the dynamics of emotional human speech also underlie emotional gait (walking). We investigate the possibility of combining these two modalities via perceptually common parameters: Speed, Intensity, irRegularity, and Extent (SIRE). We map low-level features to this 4D cross-modal emotion space and train a Gaussian Mixture Model using independent samples from both voice and gait. Our results show that a single, modality-mixed trained model can perform emotion recognition for both modalities. Most interestingly, recognition of emotion in gait using a model trained uniquely on speech data gives comparable results to a model trained on gait data alone, providing evidence for a common underlying model for emotion across modalities.

Keywords: robot emotions, emotional gait, emotional voice, affect recognition.

1 Introduction

Emotions can be conveyed in many ways outside of facial expression. Consider the sympathy we feel for a quivering puppy – he looks scared, we might say. Or the shouts of neighbors fighting in a foreign language; they can still sound angry even without knowing what they are saying. Even a singer on stage can belt out a tune with such emotional intensity that listeners are moved to tears. It is a curious phenomenon: how can mere movements or sounds affect us in this way? This kind of 'emotional intelligence' – to sense emotions through various means – appears to be built into any normal-functioning human and even some animals. We propose that robots, too, can be given the ability to understand emotions, no matter the communication channel. The goal of our research is to investigate a single model for emotion recognition, whether the channel is movement, voice, or any other type of sound.

First, consider that any movement can be colored with emotion. In the 1980's, the neurologist Manfred Clynes performed extensive cross-cultural studies using

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 52–64, 2012.

his sentograph, a device to measure touch [1]. He asked subjects to tap the device at regular intervals while imagining emotions such as love, hate, and grief. The resulting dynamic forms of the movements appear similar across cultures, e.g., abrupt, jabbing movements for hate, and soft, lethargic taps for sadness. More recently, psychologists show the importance of movement by attaching balls of light to actors' joints, turning off the lights, and recording these so-called 'point-light' displays. Actors in [2] made "drinking and knocking" movements in 10 different emotions, and despite the impoverished format, raters could still recognize emotional information. Walking style, or gait, can also reveal the walker's emotional state [19] [20]. For instance, heavyfootedness can signify anger, and slow walking speed can signify grief. For a given emotion, the dynamics of tapping, knocking, and walking already appear to have underlying similarities.

Another common way we express emotions is through the voice. In a typical study on emotional voice, researchers ask actors to utter gibberish words in various emotions. Van Bezooijen et al. [3] asked native Dutch speakers to say *twee maanden zwanger* ("two months pregnant") in neutral and nine other emotions, and then played them to Dutch and Japanese subjects. Changes in properties like pitch, tempo and loudness of speech due to physiological changes appear to create universally perceptible emotional differences [6]. Juslin and Laukka [4] reviewed dozens of studies of this kind, and found that hearers can judge anger, fear, happiness, sadness and tenderness in voice almost as well as facial expressions, around 70%. Emotion in sounds may even stretch to the animal kingdom; among some animals, alarm calls mimic human fear vocalizations, with high-pitches and abrupt onset times [7]. In primates, dominant males often emit threatening vocalizations with characteristics similar to those of human anger.

It has long been speculated that whether it be a step, tone of voice, or even a musical phrase, the expression of emotions have the same underlying 'code' [1] [4] [5]. For example, both loud, intense voices and large, forceful movements convey anger. Sadness can be conveyed through small and slow movements and quiet, slow speech. If emotions in various modalities truly share the same underlying model, then this may serve as a *common base* to combine disparate robot emotion systems instead of one model for vision, one for sound, and so on. Furthermore, it may give insight into human's ability to generalize to new situations. For example, what exactly do infants glean about emotions from their mother's voice? Can they store their knowledge about joyful speech and apply it to evaluate happy music or dance? As Breazeal states, "robots can be used as experimental testbeds for scientific investigation" to help us understand ourselves as humans [8].

In our previous studies [9] [10], we proposed a modality-independent emotion model in four dimensions representing speed, intensity, regularity and extent (SIRE)[1]. We tested this model by using SIRE features from emotional human voice to control a gesturing and music-playing robot. Human evaluators then

[1] In this paper, we change the "R" from regularity to stand for irRegularity. Whereas the perceptual feature remains the same, irregularity is simpler to represent, because it can be written in terms of the variance from a mean.

selected which emotion they perceived in the resulting robot gestures or musical sounds. These perceptual experiments showed that some combinations of SIRE invariably produced the same emotion across modalities; for example, fast, intense, irregular speech with a small pitch range produced fast, intense, irregular, small movements perceived as expressing "fear". These initial SIRE studies showed promising results, but conclusions were limited. Only a small number of samples were used, and ad-hoc scaling methods were used, i.e., "fast" and "slow" were not defined in a quantitative manner. Additionally, only emotion transfer was performed, not automatic recognition of emotion.

In this paper, we extend our SIRE approach in three ways: 1) *quantitative analysis* and scaling, made possible by a large sample size, 2) extension to analysis of *emotional gait* and 3) *machine learning* and classification. We ask the following research question: Is it possible to train a classifier using voice data, then use that classifier to recognize emotion in gait? If so, this provides evidence towards a unified emotion system, as opposed to one recognition module for each modality.

2 Approach

We investigate emotion recognition in voice and gait using four steps:

1. **Feature selection.** We select and extract low-level, modality-specific features representing Speed, Intensity, irRegularity, and Extent (SIRE). For example, *speech rate* in syllables per second is an indicator of speed in speech.
2. **Mapping to SIRE space.** We use a Gaussian normalization scheme to scale each of the four low-level features to [0,1].
3. **Personalization.** We normalize the data depending on each individual's mean speed, intensity, etc. This takes into account that elderly individuals may walk slower than average, for instance.
4. **Training** and testing in SIRE space.

2.1 Feature Selection

The general idea is to select features that may perceptually be mapped to speed, intensity, irregularity and extent. These are dynamic features that are found as principal characteristics in emotion studies across voice [11] [12], music [13] [14], and motion [15] [16] [17]. In our approach, selecting exactly which low-level feature to use is an important step up to the system designer. For example, what is a low-level definition of *extent* for voice? Both volume and pitch range are important features in vocal expressions of emotion. For the purposes of this experiment, we selected the features in Table 1 to map voice and gait to SIRE parameters.

Speech Analysis. We consider a database containing emotional speech data. Such a dataset may contain men and women of various ages saying the same

Table 1. Low-level feature to SIRE mappings

Voice feature	Parameter	Gait feature
Speech rate (syllables/sec)	Speed	Walking speed (steps/min)
Voice onset rapidity (dB/sec^2)	Intensity	Maximum foot acceleration (cm/sec^2)
Jitter (dB/sample)	irRegularity	Step timing variance (sec)
Pitch range (Hz)	Extent	Maximum step length (m)

sentence in different emotional styles, such as joyful, sad or angry. In Section 3, we describe our experiments with the Berlin Emotional Database[2], though many others such as [18] exist. Extraction of the four voice features in Table 1 has previously been described in [10].

Gait Feature Extraction. Gait studies such as [19] analyze data from multiple participants walking in various emotional styles. They may take into account walker's posture, arm swing, speed, and may use measurement instruments such as force pads, motion capture, or a combination of both: Montepare [20] and Janssen [21] considered the force of the steps, and Unuma et al. [22] took into account step-length and hip position. Montepare [23] also found correlations between emotions and perceptual cues such as smooth-jerky, stiff-loose, expanded-contracted, and so on.

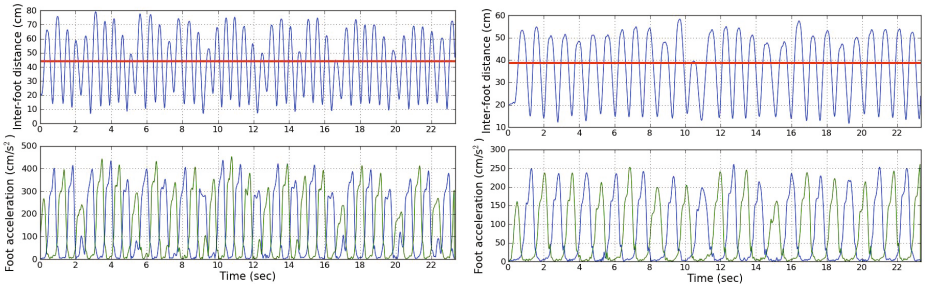
What is the minimum amount of information needed to deduce emotion in a walk? Often, gait data is collected as full-body motion capture data. For instance, in [24], thirty-five points on the body were recorded, and 30 principal components used for training and classification. This produced a recognition accuracy of 69% over five emotions. On the other hand, in a real-world human-robot interaction, it may be difficult to obtain a continuous stream of full body configuration data.

We suggest that only positions of feet through time are sufficient. This could be obtained by sensors in a human partner's shoes or slippers, for example. Our current study uses the Body Movement Library [25], which contains emotional walking by non-professionals, in neutral, happy, sad, angry, and a few samples of afraid. We use the data points of the ankle joints in x, y, z space, where z is the vertical axis.

Speed. We calculate speed in steps per minute. We subtract the position of one foot from the other in the horizontal (x, y) plane. We then perform peak picking (using average foot distance as the threshold), assuming that feet are at their maximum horizontal distance when stepping. These centroids of these peaks determine the time of each step.

Intensity. Given our dataset, we calculate the maximum acceleration achieved in the sample in x, y, z space. In a real-time situation, this may need to be used in conjunction with a sliding window. Intuitively, this corresponds to the "heavy-footedness" of the steps. In [26]'s emotion recognition approach for knocking

[2] `http://pascal.kgw.tu-berlin.de/emodb/`



(a) Example of angry gait of user "ale" (b) Example of sad gait of user "ale"

Fig. 1. Examples of gait analysis. The horizontal line indicates the threshold for peak-picking (mean value). For sad gaits, the step lengths (inter-foot distances) are shorter, and foot acceleration is lower.

movements, average acceleration was used. It is not clear whether one formulation over the other offers any advantage.

Irregularity. Step timing variance is calculated as the standard deviation in the step lengths, in seconds. For instance, walking with a "regular" pace may give a different impression compared to an "irregular" pacing which stops and starts.

Extent. This is defined as the maximum step length in x, y space.

2.2 Mapping to SIRE Space

In previous work, we performed a simple linear scaling of data between 0 and 1, relative to the maximum and minimum values of the data [10]. Here, we propose a bidirectional mapping scheme based on dataset statistical characteristics, to deduce actual speed from and S on $[0, 1]$ and vice versa.

Here, we model each of the features (speed, intensity, irregularity, extent) as Gaussian normal distributions, based on the μ and σ of the datasets (cf. Table 2 and Table 3). To perform mapping, after discretizing the feature value range, we assume the area under the curve of each segment is proportional to the amount allotted in the SIRE [0-1] range (Figure 2). This ensures more detailed sampling in the feature value ranges where samples occur most. Although we assume a single-Gaussian probability distribution function (pdf) for now, a mixture of Gaussians may prove a more accurate model. This should be tested in future work.

Table 2. Feature distributions of the Berlin Emotional Database dataset

Feature	μ	σ
Speech rate (syll/sec)	5.87	1.26
Voice onset rapidity (dB/sample^2)	11.36	4.56
Jitter (dB/sample)	871.91	269.39
Pitch range (Hz)	111.57	44.04

Table 3. Feature distributions of the Body Movement Library walking dataset

Feature	μ	σ
Walking speed (steps/min)	91.75	16.76
Maximum foot acceleration (cm/sec^2)	341.22	68.88
Step timing variance (sec)	0.07	0.06
Maximum step length (cm)	63.21	8.08

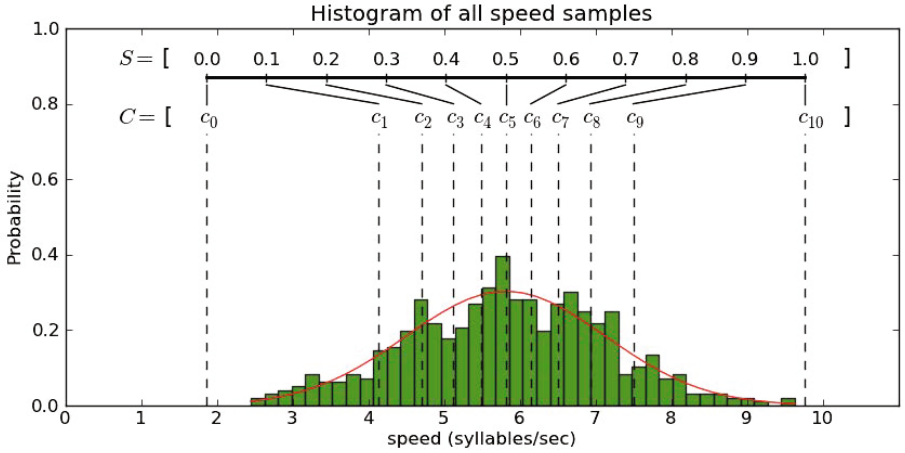


Fig. 2. Example mapping speech rate to SIRE speed using pdf fitted to data

Determining Real-to-SIRE Mapping Array C. Concretely, given μ, σ, we seek $C = \{c_0, ...c_k\}$ where $c_0 = \mu - 3 * \sigma$ and $c_{10} = \mu + 3 * \sigma$, and we numerically calculate C(k) where $k = 0, ...9$, such that

$$0.1 = cdf(x_{k+1}) - cdf(x_k). \tag{1}$$

In other words, we first define all values less or more than 3 standard deviations from the mean as 0 and 1 respectively. Then, C is a monotonically increasing sequence which defines boundaries for 10% slices of our distribution. We use these boundaries to compose a piece-wise function containing linear functions. Using C, we can define our mappings as follows. We use the example of speech and speed S here, but scaling is similarly defined for intensity I, irregularity R, and extent E using the low-level features in Table 1.

Converting SIRE Value to Real-World Value. Given C and an arbitrary value for S on $[0, 1]$, we can find a corresponding *speechRate* based on our dataset:

$$speechRate(S) = \begin{cases} c_0 + S * (c_1 - c_0) & : 0 \leq S < 0.1 \\ c_1 + (S - 0.1)(c_2 - c_1) & : 0.1 \leq S < 0.2 \\ ... & \\ c_9 + (S - 0.9)(c_{10} - c_9) & : 0.9 \leq S \leq 1.0 \end{cases}$$

Converting Real-World Value to SIRE Value. Similarly, to find S, the speed value, we can define a mapping given C and a real-world value x:

$$S(x) = \begin{cases} 0 & : x < c_0 \\ 0.1(x - c_0)/(c_1 - c_0) & : c_0 \leq x < c_1 \\ 0.1 + 0.1(x - c_1)/(c_2 - c_1) & : c_1 \leq x < c_2 \\ ... & \\ 0.9 + 0.1(x - c_9)/(c_{10} - c_9) & : c_9 \leq x \\ 1 & : c_{10} < x \end{cases}$$

Both Equations can be similarly defined for I, R, E.

2.3 Personalization and Training

Following the result of [26], we adjust each sample depending on the relative difference of the individual compared to the dataset average. Given k emotional samples created by an individual $P = p_0, .., p_k$, and a given feature $f \in S, I, R, E$, we can find the individual's average speed $\mu(P_S)$, intensity $\mu(P_I)$, and so on. Then, given a group (dataset) G of n individuals, we can also determine the group's average feature values $\mu(G_f)$. We determine a personalized bias $b(P_f)$ for each individual/feature pair in the dataset, and define it as $b(P_f) = \mu(G_f) - \mu(P_f)$. In the personalization step, we update each sample: $P_{f,k} = P_{f,k} + b(P_f)$.

Given the datasets described above, we perform training using expectation maximization to train a K-mixture Gaussian Mixture model, where K is the number of classes (e.g. happy, sad, angry, scared, and neutral). It possible that we would need more components to model each emotion. For example, a "fear" gait may have two possibilities for speed: slow for a hesitant approach, or fast to run away. The number of mixtures should be explored in future work.

Multi-modal Training. Since the two modalities are mapped to the same perceptual feature set, we can propose a new multi-modal recognition scheme which differs from traditional multi-modal fusion approaches. In Figure 3, we can see that the SIRE classification scheme is simple, if we can assume that both modality sources can independently add information to generalize about the other. In this example, we see how it assures low dimensionality compared to the concatenation method, in which classifiers train on a feature vector containing values from multiple sources. Indeed, in (c), the weights also add dimensions for optimization.

3 Experiments

3.1 Purpose

We aim to answer the following research questions: 1) What are the real-world values defining emotions in speech and gait? 2) Are the SIRE values defining emotions in speech and gait similar? 3) What is the effect of using SIRE mapping

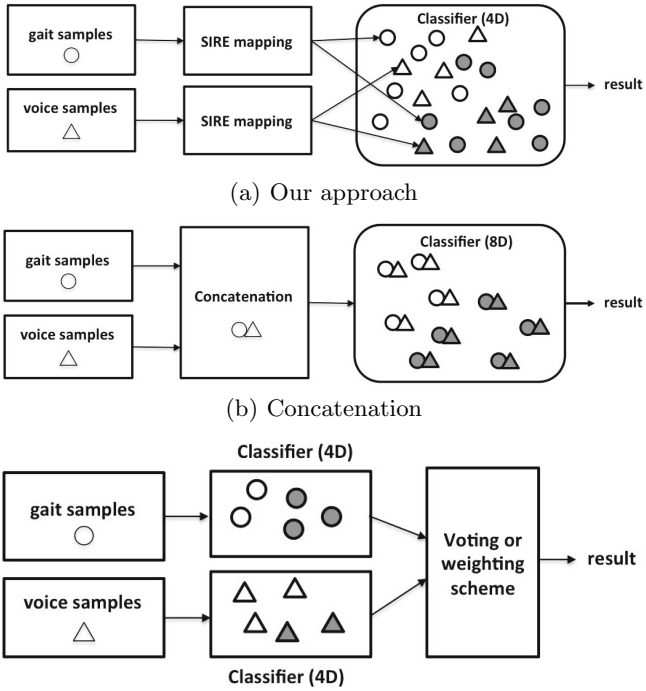


Fig. 3. Comparison of approaches for integration and classification of multimodal data

and personalization on emotion training and recognition? 4) Can modalities be integrated using the training scheme described in Sec. 2.3? What is the effect? 5) Can emotion classifiers be trained with one modality and tested with another?

3.2 Materials and Procedure

For the emotional walking data, we used 28 subjects from the Body Movement Library [25]. Each individual provided 2 samples each of happiness, sadness, anger and neutral styles of walking. Six subjects also provided 2 samples each of fear. In total, 236 gait samples were used. Each sample was approximately 23 seconds long.

For emotional speech data, we used 10 subjects (5 female, 5 male) from the Berlin Emotional Speech database, who spoke up to 10 different sentences in happy, sad, angry, neutral and fear styles. Some were not included in the analysis if the recognition rate by humans was not at least 80%. In total, we used 408 voice samples from this database.

We use the Scikit-learn [27] toolkit to train a 5-component Gaussian mixture model and perform a recognition step. We perform 10-fold cross validation with the following training and testing sets: 1) voice samples only 2) gait samples only 3) voice and gait samples 4) training with voice and testing with gait, and 5) training with gait and testing with voice.

Table 4. Mean real-world values for emotions based on voice samples

Feature	Speech rate (syll/sec)	Voice onset rapidity (dB/sample^2)	Jitter (dB/sample)	Pitch range (Hz)
Happiness	6.1	13.0	871	144
Sadness	4.3	8.5	724	101
Anger	6.0	13.7	964	131
Fear	7	10.8	1025	105
Neutral	6.4	10.3	754	82

Table 5. Mean real-world values for emotions based on gait samples

Feature	Walking speed (steps/min)	Acceleration (cm/s^2)	Variance (ms)	Step length (cm)
Happiness	96	362	64	65
Sadness	76	272	77	56
Anger	105	411	63	71
Fear	92	324	78	62
Neutral	90	323	58	61

4 Results and Discussion

4.1 Quantitative Descriptions of Emotions

Real-world values defining emotions in speech and gait, in terms of our feature set, are given in Table 4 and 5. These values can be used for generation of emotionally expressive robot voices or gaits.

Our most useful results come from comparing emotional styles to neutral. For instance, a fear gait is very similar to a neutral gait, except that step timing is more irregular. An angry gait has much higher values on all dimensions compared to neutral, except that step timing is relatively regular compared to other emotions. Happiness is conveyed with slightly higher values along all dimensions. Sadness values are lower than neutral on all dimensions, except for step timing: steps are irregular – perhaps indicating a faltering, energy-less gait.

In future work, it will also be interesting to compare these values to real-world values of SIRE in music. For example, the *adagio* tempo typically found in sad songs is defined for classical music at around 66-76 beats per minute. Here, we also find "sad" gait walking speed at 76 steps per minute. There may be absolute thresholds defining emotions in a general sense. This would be an interesting direction for investigation.

4.2 SIRE Descriptions of Emotions

In our previous work [9], we published the perceptually-scaled SIRE parameters of the highest rated sample for a given vocal emotion. Now we use the information

Table 6. Mean SIRE values for emotions based on voice and gait samples (S=speed, I=intensity, R=irregularity, E=extent)

Voice	S	I	R	E
Happiness	0.59	0.63	0.49	0.74
Sadness	0.13	0.27	**0.29**	**0.40**
Anger	**0.56**	0.68	0.62	0.65
Fear	**0.81**	0.45	0.70	0.43
Neutral	0.66	0.41	0.34	0.25

Gait	S	I	R	E
Happiness	0.60	0.61	0.49	0.64
Sadness	0.18	0.16	**0.58**	**0.19**
Anger	**0.78**	0.84	0.48	0.83
Fear	**0.51**	0.41	0.58	0.39
Neutral	0.46	0.41	0.44	0.39

from many samples and compare the voice results to gait data. In comparing the two modalities in Table 6, we see many SIRE values overlap across modalities. For instance, the speed for happiness is about 0.6 for both voice and gait. We highlight those which differ more than 15%.

We can see that sadness in voice was found to be fairly regular, and in motion more irregular. Additionally, sadness "extent" is much smaller in gait than voice. One possible explanation is the interpretation of "sadness". Consider that Parrot's emotion hierarchy [28], further classifies "sadness" into more specific terms, such as "anguish" and "depression". Anguish may manifest itself with a pleading, highly variant pitch, versus a more monotone depressed style. This indicates the necessity to specify subcategories of emotion in future studies. Similarly, the voice database specified that anger was "hot anger", whereas the individuals in the gait database were free to choose their interpretation of anger.

As for the stark difference in speed for fear, we suggest that speed appears to fall into two categories – slow and hesitant (fear of approaching something) and fast (running away from something). According to the Body Movement library authors [25], participants expressed difficulty with this emotion because they felt the need to express fear relative to an object. This indicates the importance of the object in question and its orientation relative to the robot in expressing the fear emotion correctly.

4.3 Recognition Results

Here, we look at the effect of using SIRE mapping and personalization on emotion training and recognition. In Figure 4, we can see that the personalized SIRE method gives the best results in all cases, given chance level at 20%. Personalization seemed particularly important in the gait dataset. This may be due to the fact that the voice dataset was created using actors, who may have learned a standard method to convey emotions.

In terms of multimodal training, we can see in Figure 4(c), that SIRE allows for integration of the two features, where a simple bag of features approach fails. On the other hand, using the augmented voice+gait dataset does not appear to show any significant improvement over training with voice or gait respectively.

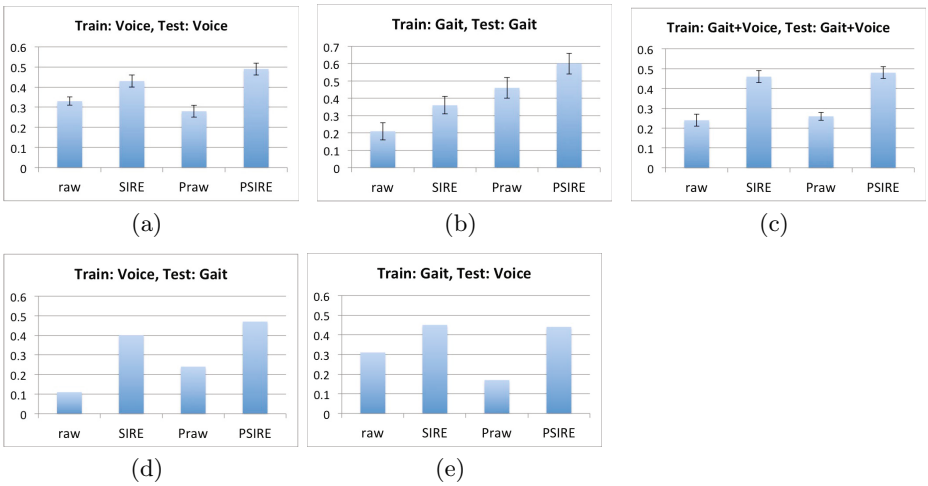


Fig. 4. Recognition results using different methods (raw=real-world features, SIRE=SIRE features, Praw=personalized raw features; PSIRE=personalized SIRE features) and test-train combinations

Most interestingly, cross-modal learning was observed. In Figure 4(d) and (e), we see that it is indeed possible to train with one modality and test on another. SIRE mapping is essential, and gives 46% accuracy testing on gait when the model was trained with voice.

5 Conclusion

In this paper, we proposed a new approach to detecting emotion in gait for robot understanding of emotion. Along the way, we answered the following research questions, giving evidence to support our SIRE model of emotion:

1. *What are the real-world (quantitative) dynamic values defining emotions in speech and gait?* These values are given in Table 4 and 5.
2. *What are the SIRE values defining emotions in speech and gait, and are they similar?* Yes, they are similar to a certain extent. Differences may be attributed to varying interpretations of emotions, such as "anguish" vs. "depressed" sadness.
3. *What is the effect of using SIRE mapping and personalization on emotion training and recognition?* Personalization and SIRE mapping together provide the best performance in our independent classification tests.
4. *Can modalities be integrated using the training scheme described in Sec. 2.3? What is the effect?* SIRE allows the integration of the voice and gait modalities in the same space, giving comparable results to separate recognition modules. This suggests that it is possible to model emotion (of a robot, for example) in one unified space.

5. *Can an emotion classifier be trained with one modality and tested with another?* Yes, in training with voice and testing with gait, we showed up to 46% recognition compared to a chance baseline of 20%.

Although the recognition result rate may not be as high as other methods using higher dimensional feature sets, this study provides additional evidence to an underlying 4-parameter emotion model across voice and gait. In particular, values for emotions in SIRE space for voice and gait are so similar that training with one modality allows recognition in the other. Future work should include examining the improvement of recognition when adding other cues (e.g. head down versus head up), taking into account emotional intensity (e.g. somewhat angry versus very angry), and evaluation of the SIRE model in a human-robot interaction setting.

References

1. Clynes, M.: Sentics: The Touch of the Emotions. Prism Press, UK (1989)
2. Pollick, F.E., Paterson, H.M., Bruderlin, A., Sanford, A.J.: Perceiving affect from arm movement. J. Personal. 82, 51–61 (2001)
3. Van Bezooijen, R., Van Otto, S.A., Heenan, T.A.: Recognition of vocal dimensions of emotion: A three-nation study to identify universal characteristics. J. Cross-Cultural Psych. 14, 387–406 (1983)
4. Juslin, P.N., Laukka, P.: Communication of emotions in vocal expression and music performance: different channels, same code? Psychol. Bull. 129(5), 770–814 (2003)
5. Spencer, H.: The origin and function of music. Fraser's Magazine 56, 396–408 (1857)
6. Scherer, K.H.: Vocal affect expression: A review and a model for future research. Psychol. Bull. 99, 143–165 (1986)
7. Snowdon, C.T.: Expression of emotion in non-human animals. In: Davidson, R.J., Sherer, K.H., Goldsmith, H.H. (eds.) Handbook of affective sciences, pp. 457–480. Oxford University Press, London (2003)
8. Breazeal, C.: Designing sociable robots, 1st edn. The MIT Press, Cambridge (2004)
9. Lim, A., Ogata, T., Okuno, H.G.: Towards expressive musical robots: a cross-modal framework for emotional gesture, voice and music. EURASIP J. Audio, Speech, and Music Proc. 2012(3) (2012)
10. Lim, A., Ogata, T., Okuno, H.G.: Converting emotional voice to motion for robot telepresence. In: Humanoids, Bled, pp. 472–479 (2011)
11. Cowie, R., et al.: Emotion recognition in human-computer interaction. IEEE Signal Proc. Magazine 18(1), 32–80 (2001)
12. Fernandez, R., Picard, R.W.: Classical and Novel Discriminant Features for Affect Recognition from Speech. In: INTERSPEECH, pp. 4–8 (2005)
13. Mion, L., De Poli, G.: Score-independent audio features for description of music expression. IEEE Trans. Audio Speech Lang. Process. 16(2), 458–466 (2008)
14. Livingstone, S.R., Brown, A.R., Muhlberger, R., Thompson, W.F.: Modifying score and performance changing musical emotion: a computational rule system for modifying score and performance. Comput. Music J. 34(1), 41–65 (2010)
15. Amaya, K., Bruderlin, A., Calvert, T.: Emotion from motion. Graph. In: Interface, pp. 222–229 (1996)
16. Pelachaud, C.: Studies on gesture expressivity for a virtual agent. Speech Commun. 51(7), 630–639 (2009)

17. Camurri, A., Volpe, G.: Communicating expressiveness and affect in multimodal interactive systems. Multimedia 12(1), 43–53 (2005)
18. Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., Karpouzis, K.: The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 488–500. Springer, Heidelberg (2007)
19. Roether, C.L., Omlor, L., Christensen, A., Giese, M.A.: Critical features for the perception of emotion from gait. J. Vision 9(6), 15, 1–32 (2009)
20. Montepare, J.M., Goldstein, S.B.: The identification of emotions from gait information. J. Nonverbal Behav. 11(1), 33–42 (1987)
21. Janssen, D., et al.: Recognition of emotions in gait patterns by means of artificial neural nets. J. Nonverbal Behav. 32, 79–92 (2008)
22. Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. In: SIGGRAPH, Los Angeles, pp. 91–96 (1995)
23. Montepare, J., Koff, E., Zaichik, D., Albert, M.: The use of body movements and gestures as cues to emotions in younger and older adults. J. Nonverbal Behav. 23(2), 133–152 (1999)
24. Karg, M., Kuhnlenz, K., Buss, M.: Recognition of affect based on gait patterns. IEEE Trans. Sys., Man, Cyber. 40(4), 1050–1061 (2010)
25. Ma, Y., Paterson, H.M., Pollick, F.E.: A motion-capture library for the study of identity, gender, and emotion perception from biological motion. Behav. Res. Meth., Inst., & Comp. 38, 134–141 (2006)
26. Bernhardt, D.: Detecting emotions from everyday body movements. Presenccia PhD Sym., Barcelona (2007)
27. Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011)
28. Parrot, W.G.: Emotions in social psychology. Philadelphia Press, Philadelphia (2001)

Gender Differences in the Perception of Affective Movements

Ali-Akbar Samadani, Rob Gorbet, and Dana Kulić

University of Waterloo, Electrical and Computer Engineering Department
Waterloo, Ontario N2L 3G1, Canada

Abstract. Identifying human capabilities in perceiving affective expressions is essential for developing interactive machines that can engage with their human users. In order to ensure that the behaviour of the interactive machine is perceived as intended, any gender–specific differences in the perception of affective expressions are an important design consideration. This paper presents a preliminary study investigating the role of gender in the perception of affective hand movements displayed on both anthropomorphic and non-anthropomorphic structures. The results show that gender significantly influences the participants' perception and that the impact of the display structure and intended-emotion on the perception of the affective movements differs between male and female observers.

Keywords: Affective movements, Gender differences, Display structure, Perception, User study.

1 Introduction

Humans associate different body movements and postures with distinct affective expressions (e.g., anger is associated with frequent tempo changes) [6], [34], [12], [20], [1], and are able to identify the feeling encoded in a displayed movement even when demonstrators try to conceal their expression (e.g., negative body language) [4], [22], [10], [8]. Moreover, the psychology literature reports on the human tendency to ascribe human–like social and affective attributes to non–anthropomorphic structures such as abstract moving geometrical shapes, and even consider them to be engaging in social interactions [15]. Affective movement recognition and generation capabilities are particularly important in the field of human–machine interaction, in applications such as robotic social agents, kinetic sculptures, and animated characters. In order to develop reliable computational models for automatic affective movement recognition and generation for autonomous systems, it is important to understand how humans perceive affect from movement and whether there are gender-specific differences in the perception of affective movement.

The present work is a collaboration with Philip Beesley Architect Inc., a design practice developing a series of architectural responsive environments, called the *Hylozoic* series [3], [2]. These environments use massively repeating components, microprocessors, sensors and actuators to create decentralized responsive

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 65–76, 2012.

systems capable of subtle motions giving the impression that the environments are 'sensitive' and may even have affective states (Figure 1). The long term goal of our research is to develop sufficient understanding of affective movement generation and perception to enable these structures to engage in affective communication with their occupants through movement.

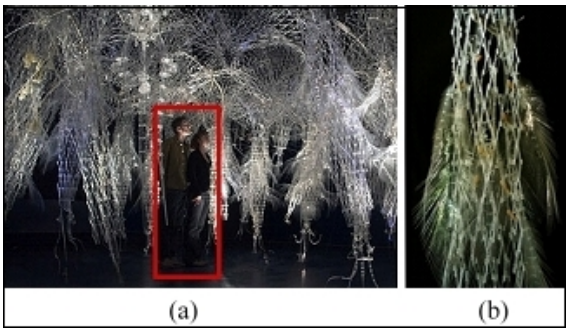


Fig. 1. a) Two visitors highlighted with a red outlined rectangle immersed in Hylozoic Soil, a responsive architectural geotextile environment [2]. b) Hylozoic Soil consists of layers of mechanical fronds and whiskers that move in response to the human occupants [2]. Reprinted with permission.

The effect of gender on the perception of body language and in particular, bodily expression of emotion is largely unexplored. Differences in affective movement perception could arise due to the gender of the demonstrator and/or observer. Furthermore, the structure on which the affective movement is displayed may have a different effect on how the emotion is perceived by male or female participants. In an early study by Carmichael et al. [7], behavioural hand and arm gestures performed by an actor (e.g., hand and arm gestures for prayer, fear, anxiety) were correctly recognized above the chance level and no significant gender-differences in the perception of the gestures was observed. In general, reports on gender differences in the perception of affective expressions mainly focus on facial expressions. Male and female high school, college and university students showed significant differences in their rating of facial expressions corresponding to the six Ekman emotions [17]. Women perceive conveyed emotions through facial expressions more accurately than men [17, 13]. In another study, participants were shown videos of neutral faces gradually changing to express different emotions and women were more accurate and sensitive in perceiving the facial expressions [24]. Furthermore, neurological studies report on the involvement of different underlying circuitry in perception of emotion in men and women [32].

Other studies investigate the role of the demonstrator's gender in the perception of affective movements. In a user study, participants tended to apply social stereotypes to infer the gender of a point-light display throwing a ball with different emotions: happiness, sadness, anger, and neutral. Angry movements were

perceived to be demonstrated by men and sad movements were more likely to be attributed to women [16]. Due to kinematic similarities between fearful gait and female gait, the fearful gait is better perceived if the walker is female [14]. Significant gender differences in the perception of emotion from static postures of *Venus* and *Apollo* with different arm positions are reported in [30].

According to these studies, gender might play an important role in the perception of affective movement; hence, further investigation is needed to identify the role of gender in affective movement perception. To the best of the our knowledge, there has been no research reported on gender differences in the role of display structure on the perception of affective movements.

We have conducted a user-study in which participants watched videos of a set of affective hand movements displayed on human-like and frond-like structures (Figure 2) and evaluated the perceived affective expressions. The frond-like structure appearance was designed to be similar to the Hylozoic soil structural elements. In a previous study, the effect of the intended–emotion and display structure on the participants' perception of the movements was investigated and it was found that the intended–emotion has a main affect on the participants' perception of the affective movements and that the participants' perception of the affective movements was significantly affected by the display structure, specifically in the case of sad movements [28]. In the present study, we investigate the following questions:

1. Did the gender of the observers have an influence on the perception of affective hand movements?
2. Did the intended–emotion and display structure have a different impact on male or female observers?

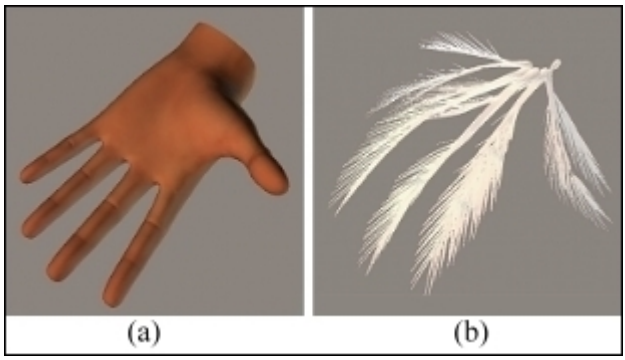


Fig. 2. Structures used to display expressive movements. a) anthropomorphic (human–like) hand model, b) non–anthropomorphic frond–like structure. These animated structures are produced using Poser (version 8, Smith Micro Inc.).

2 Affective Human Hand Movements

The labeled dataset from [29] is used in this study, and includes one movement type, closing and opening of the hand, which mainly involves phalangeal and carpo–metacarpal joint movements. Three different affective expressions were considered: sadness, happiness and anger. Five repetitions of each expression were collected. A demonstrator, who has been exposed to Laban notation [18], and is familiar with other human movement perception works (e.g. Camurri et al [6]), performed the hand movements while wearing a data glove (ShapeHand from Measurand [21], [23]). Videos of these movements are available in [27].

The movements were animated on each of the two structures shown in Figure 2. These structures have the same kinematics but their physical appearance differs. The rationale for choosing hand movements in this study is that the hand is an important medium for communicative gestures [34], and it closely resembles the motion style and structure of the moving components of the Hylozoic environments.

3 Questionnaire Study

In order to assess how affective movements are perceived and any impact of display structure, human observers were asked to rate the level of observed affective expression in each movement.

Observers were asked to rate the level of affect using both a discrete and dimensional emotion models. For the discrete model, the well known Ekman model was used, which proposes anger, happiness, sadness, surprise, disgust and fear as the six basic and universally recognized emotions [11]. For the dimensional model, the Circumplex model of emotion [26] was used, which represents emotions in a continuous two dimensional space defined by arousal and valence. The arousal dimension represents the intensity of an emotion and the valence dimension ranges from negative (unpleasant) to positive (pleasant).

During the user study, videos of the movements performed on the two different structures (Figure 2) were shown to the participants. They were then asked to evaluate the demonstrated movements in terms of expressivity. A total of 22 participants (26.1 years $\pm$ 5.8 years, 12 male, 10 female) completed the questionnaire. Participants were healthy adults with a basic working knowledge of computers and were students at the University of Waterloo. They were provided with detailed information on the study and the procedure to complete the computer–based questionnaire. All the questionnaire sessions took place at the same location and were administered by the author to ensure a uniform experience for all the participants. The study received ethics approval from the Office of Research Ethics, University of Waterloo, and a consent form was signed electronically by each participant prior to the start of the questionnaire.

In a questionnaire session, a participant watches and rates the same affective movements displayed on the human–like hand structure (Figure 2.a) and the frond–like structure (Figure 2.b). The following naming format is used to refer

to the animations in the rest of the paper: "(structure: hand, frond)_(intended–emotion: angry, happy, sad)" (e.g., "Hand_Happy" represents the happy movement displayed on the human–like hand structure).

The animations of affective hand movements were shown to the participants in randomized order. Each video was accompanied by two questions. The first question was a multiple–selection question asking participants to select among a list of keywords those that most closely described the animated structure in the video. Detailed data analysis for the first question can be found in [28].

The second question asked the participants to rate on a Likert scale the extent to which each of the six Ekman basic emotions was conveyed in the displayed animation, with 1 being "not conveyed at all" and 6 being "strongly conveyed". We used all six Ekman emotions in the questionnaire to determine emotion recognition capabilities accurately. Offering participants the choice of six emotions gives a more accurate picture of recognition rate, since it does not artificially constrain the responses and shows whether emotions are unambiguously recognized. In the third question, participants were asked to rate the arousal and valence components of the emotion perceived for each displayed movement, using a 7-point scale. A brief description of the arousal and valence dimensions of emotion was provided, along with a schematic representation of Circumplex model of emotion adapted from [9] and shown in Figure 3. Low intensity–high intensity and unpleasant–pleasant are the adjective pairs displayed at the extremes of arousal and valence scales, respectively, to further guide the participants in evaluating the arousal and valence components.

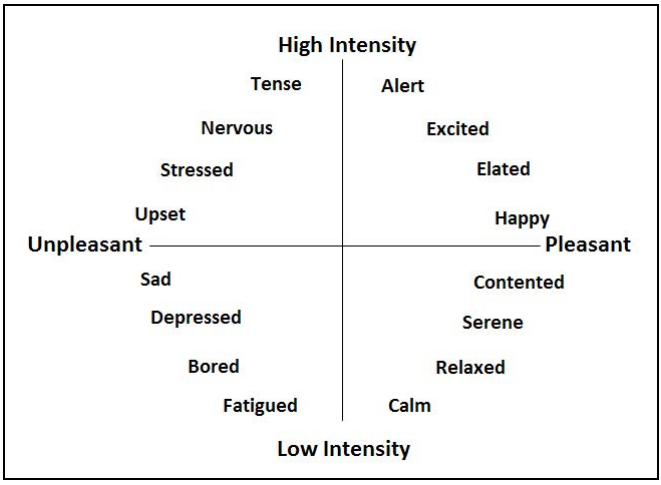


Fig. 3. A schematic representation of affective Circumplex used in the questionnaire. This figure is adapted from [9].

4 Questionnaire Data Analysis

To investigate the effect of gender and its interaction with the display structure and intended–emotion on the participants' ratings of the affective movements, a three-way repeated measure ANOVA can be used with gender as a between-group variable, and display structure and intended–emotion as within-group variables. However, the interpretation of the significant effects from a three-way repeated measure ANOVA is difficult due to the large number of variables and their main and interaction effects (7 main and interaction effects). Furthermore, a larger sample size would be needed to detect significant effects of all the variables presented in the study. To reduce the number of effects and simplify the analysis, we instead performed two sets of two-way repeated measure ANOVAs (each set contains five ANOVA tests) to assess the main and interaction effects of the intended–emotion and structure on the ratings of anger, happiness, sadness, arousal, and valence by the male participants (set 1) and female participants (set 2). This way, we have reduced the number of variables to two within-group variables: intended–emotion and display structure. Therefore, the number of main and interaction effects is reduced to three, which facilitates the interpretation of different effects on the male and female participants' perception. Table 1 shows the null hypotheses tested in each repeated measure ANOVA.

Tables 2 and 3 show the resulting F–statistics, p–values, and effect sizes (η^2) for male and female participants, respectively. The SPSS statistical software package [31] was used to generate the user study results. The ANOVA results are considered significant at $p<0.05$.

According to the ANOVA results in Table 2, there is a significant interaction between structure and intended–emotion in the male participants' ratings of anger, happiness, sadness, and valence; hence rejecting $H_0^{male}(3,i)$ for $i = \{anger, happiness, sadness, valence\}$. However, no significant interaction between the intended–emotion and structure in the female participants' ratings was observed (Table 3); hence, retaining $H_0^{female}(3,i)$'s.

There are also differences in the main effects of the intended–emotion and display structure on the male and female participants' perception. The intended–emotion was found to significantly influence the ratings of both the male and

Table 1. Null hypotheses tested in the repeated measure ANOVAs for the male participants' ratings; $i = \{Anger, Happiness, Sadness, Arousal, Valence\}$, $G = \{male, female\}$

$H_0^G(1,i)$:	*The means of the G participants' ratings of i for different intended–emotions are equal.*
$H_0^G(2,i)$:	*The means of the G participants' ratings of i for different structures are equal.*
$H_0^G(3,i)$:	*Structure and intended–emotions are independent and no interaction effect between the two is present in the G participants' ratings of i.*

Table 2. F–statistics, p-values and effect size (η^2) results from two-way repeated measure ANOVAs each testing the main and interaction effects of structure and intended–emotion on male participants' ratings of anger, happiness, sadness, arousal, and valence. There are 12 male participants. Greenhouse-Geisser correction is used when sphericity assumption is violated. "*" sign indicates a significant effect. Bonferroni adjustment was made for multiple comparisons.

	i : Anger	Happiness	Sadness	Arousal	Valence
Intended–Emotion	$F(2,22) = 15.006$	20.749	2.127	18.947	9.981
$\left(H_0^{male}(1,i)\right)$	$p = 0.000$*	0.000*	0.143	0.000*	0.001*
	$\eta^2 = 0.315$	0.244	0.089	0.503	0.162
Structure	$F(1,11) = 0.014$	4.068	3.000	1.232	2.129
$\left(H_0^{male}(2,i)\right)$	$p = 0.908$	0.069	0.111	0.291	0.172
	$\eta^2 = 0.000$	0.027	0.014	0.004	0.012
Structure x Intended–emotion	$F(2,22) = 5.421$	7.871	8.406	1.329	17.488
$\left(H_0^{male}(3,i)\right)$	$p = 0.012$*	0.003*	0.002*	0.285	0.000*
	$\eta^2 = 0.071$	0.166	0.103	0.008	0.228

Table 3. F–statistics, p-values and effect size (η^2) results from two-way repeated measure ANOVAs each testing the main and interaction effects of structure and intended–emotion on the female participants' ratings of anger, happiness, sadness, arousal, and valence. There are 10 female participants. Greenhouse-Geisser correction is used when sphericity assumption is violated. "*" sign indicates a significant effect. Bonferroni adjustment was made for multiple comparisons.

	i : Anger	Happiness	Sadness	Arousal	Valence
Intended–Emotion	$F(2,18) = 8.825$	7.676	14.333	33.081	15.221
$\left(H_0^{female}(1,i)\right)$	$p = 0.002$*	0.004*	0.000*	0.000*	0.000*
	$\eta^2 = 0.254$	0.174	0.230	0.612	0.311
Structure	$F(1,9) = 16.308$	1.385	11.000	11.184	38.383
$\left(H_0^{female}(2,i)\right)$	$p = 0.002$*	0.269	0.009*	0.009*	0.000*
	$\eta^2 = 0.114$	0.009	0.042	0.047	0.122
Structure x Intended–emotion	$F(2,18) = 1.619$	2.739	3.508	0.360	2.521
$\left(H_0^{female}(3,i)\right)$	$p = 0.226$	0.092	0.084	0.703	0.108
	$\eta^2 = 0.019$	0.050	0.064	0.003	0.060

female participants in all the cases in this study except for the male sadness ratings. The structure has a significant main effect on the female participants' ratings in all the cases at $p < 0.05$ except for the happiness ratings (rejecting $H_0^{female}(2,i)$ for $i = \{anger, sadness, arousal, valence\}$), whereas the effect of structure on the male participants' ratings was not found significant at $p < 0.05$.

Bar charts of average ratings of anger, happiness, sadness, arousal, and valence by male and female participants are shown in Figure 4. Paired t-tests are performed between the pairs of the male and female participants' ratings of the

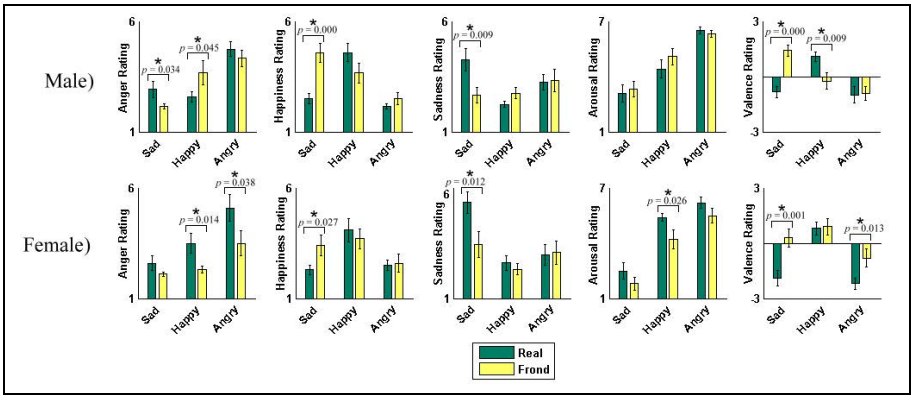


Fig. 4. Average ratings (mean ± SE) for the affective movements displayed on the human–like and frond–like structures by 12 male and 10 female participants. From left to right, ratings for: anger, happiness, sadness, arousal, valence. Significant pair-wise differences between the ratings of an intended–emotion displayed on different structures are indicated using "*" sign and their p-values are reported.

affective movements displayed on the hand–like and frond–like structures and significant pair-wise differences are shown using "*" in Figure 4.

Table 4 shows a confusion matrix of the perception of the intended–emotions (i.e., anger, happiness, sadness). For the confusion matrix, an emotion is considered recognized if it is rated 3 or above on the Likert scale. Note that this recognition cut–off is applied only for illustrative purposes in Table 3 and all the analysis in Section 4 is done on the full scale of ratings obtained in the questionnaire study.

As can be seen in Table 4, the perception of anger by the female participants was significantly affected by the structure as the angry movement displayed on the frond–like structure was less frequently recognized as conveying anger in comparison with the angry movement displayed on the human-like structure (female anger rating of the angry movements in Figure 4). The male participants equally attributed high-arousal and negative valence to the angry movement and correctly recognized angry movement regardless of the structure (male anger, arousal, and valence ratings of the angry movement in Figure 4). However, female participants associated a lower-level of arousal and less-negative valence to the frond–like structure displaying the angry movement (female arousal and valence ratings of the angry movement in Figure 4). The better performance of the male participants in recognizing angry movements is congruent with [33, 25] suggesting that men are more accurate in recognizing angry expressions. The happy movement displayed on the human–like structure is correctly recognized as conveying happiness and positive valence by both male and female participants, whereas the frond–like structure displaying happiness is less frequently recognized as happy. The male participants frequently misperceived the happy movement displayed on the frond–like structure as conveying anger, which might be the reason for the slightly negative valence attributed to the Frond_Happy

movement by the male participants. Frond_Happy movement is correctly recognized by the female participants. Although there is a significant difference between the average arousal ratings of the Frond_Happy and Real_Happy movements by the female participants, these average ratings are relatively high for both structures with the Real_Happy movement regarded as conveying a higher arousal. The relatively higher accuracy of the female participants in recognizing happy movements in comparison to the male participants is similar to the reports in [35, 5] suggesting that women are more tuned to experiencing positive expressions.

Both the male and female participants correctly rate the sad movement displayed on the human–like structure as sad with low arousal and negative valence attributes, while the Frond_Sad movement is less frequently recognized as sad. The Frond_Sad movement is frequently perceived as conveying happiness and positive valence, especially by the male participants. Overall, both male and female participants correctly recognized differing levels of arousal from the affective movements, while women rate the perceived valence more accurately, which is consistent with [19].

Another important observation in this user–study is that the male and female participants exhibit a more similar affective movement perception when the demonstrator structure is human-like (Figure 4 and Table 4). Such structure-specific similarities in the perception of affective movements merit further inves-

Table 4. Confusion matrix showing percentage (%)* of anger, happiness, and sadness ratings for different affective movements by the 12 male and 10 female participants. The recognition rates greater than 50% are highlighted.

	Perceived emotions		
	Anger	Happiness	Sadness
Hand_Angry (male)	**92%**	0%	33%
Frond_Angry (male)	**75%**	8%	33%
Hand_Angry (female)	**70%**	10%	30%
Frond_Angry (female)	30%	20%	40%
Hand_Happy (male)	17%	**83%**	0%
Frond_Happy (male)	50%	50%	25%
Hand_Happy (female)	50%	**70%**	30%
Frond_Happy (female)	0%	**60%**	20%
Hand_Sad (male)	25%	17%	**58%**
Frond_Sad (male)	0%	**67%**	17%
Hand_Sad (female)	20%	10%	**90%**
Frond_Sad (female)	0%	40%	40%

∗ There are cases where an affective movement was rated 3 or above for more than one emotion. On the other hand, there are cases in which anger, happiness and sadness were all rated below 3. This is why none of the emotion ratings add up to 100% in the confusion matrix.

tigation and would potentially motivate the use of more human–like structures for communicating affect during human-robot interaction to ensure consistent perception.

5 Conclusions

User studies allow for the exploration of the human capabilities in recognizing affective expressions displayed on different structures. Insight gained from such user studies can inform the design of interactive technologies capable of displaying various affective expressions. To the best of our knowledge, this study is the first report on gender differences in the perception of dynamic structures displaying affective movements. In the preliminary study presented in this paper, gender-specific differences in the perception of affective hand movements displayed on two different structures were investigated. It was found that the gender significantly influenced the perception of the affective movements in many cases. Furthermore, cases were observed in which the impact of the intended–emotion and display structure on the participants' perception of the affective movements varied between male and female participants (e.g., anger ratings for Frond_Angry movement). The male participants perceived angry movements more accurately than the female participants regardless of the display structure, whereas the female participants performed better in recognizing happy movements. Both male and female participants frequently misperceived sad movements displayed on the frond-like structure as conveying a positive expression.

The detected main and interaction effects of the intended–emotion and display structure in this study are of medium to large sizes. These findings demonstrate the important role that gender might play in the perception of affective movements and emphasize the importance of considering gender in the design of affective display mechanisms in general. There are a few prominent effects (e.g., intended–emotion effect on the sadness ratings of the male participants) that were not detected in this study. Future studies with a larger sample size will enable investigating the importance of these effects.

In future studies, the role of gender in the perception of affective movements will be further explored with a larger number of participants and a larger variety of affective movements in terms of expressivity and motion path. Furthermore, gender differences in the perception of different display structures will be further investigated to identify if there exist structures that might limit (or modulate) the communication of affective expression with male or female observers.

References

[1] Argyle, M.: Bodily communication. Taylor & Francis (1988)
[2] Beesley, P.: Hylozoic soil. Leonardo 42(4), 360–361 (2009)
[3] Beesley, P.: Kinetic architectures and geotextile installations. Riverside Architectural Press (2010)
[4] Blake, R., Shiffrar, M.: Perception of human motion. Annual Review of Psychology 58, 47–73 (2007)

[5] Brody, L.: Gender, emotional expression, and parent-child boundaries. Emotion: Interdisciplinary Perspectives, 139–170 (1996)
[6] Camurri, A., Lagerlöf, I., Volpe, G.: Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. International Journal of Human-Computer Studies 59(1–2), 213–225 (2003)
[7] Carmichael, L., Roberts, S.O., Wessell, N.Y.: A study of the judgment of manual expression as presented in still and motion pictures. The Journal of Social Psychology 8(1), 115–142 (1937)
[8] Castellano, G., Mancini, M., Peters, C., McOwan, P.W.: Expressive copying behavior for social agents: A perceptual analysis. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 42(3), 776–783 (2012)
[9] Colibazzi, T., Posner, J., Wang, Z., Gorman, D., Gerber, A., Yu, S., Zhu, H., Kangarlu, A., Duan, Y., Russell, J.A.: Neural systems subserving valence and arousal during the experience of induced emotions. Emotion 10(3), 377–389 (2010)
[10] Coulson, M.: Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. Journal of Nonverbal Behavior 28(2), 117–139 (2004)
[11] Ekman, P.: Are there basic emotions? Psychological Review 99(3), 550–553 (1992)
[12] Fast, J.: Body language. Pocket (1988)
[13] Hall, J., Matsumoto, D.: Gender differences in judgments of multiple emotions from facial expressions. Emotion 4(2), 201–206 (2004)
[14] Halovic, S., Kroos, C.: Facilitating the perception of anger and fear in male and female walkers. In: Proc. of AISB, Symposium on Mental States, Emotions and their Embodiment, pp. 3–7 (2009)
[15] Heider, F., Simmel, M.: An experimental study of apparent behavior. The American Journal of Psychology 57(2), 243–259 (1944)
[16] Johnson, K.L., McKay, L.S., Pollick, F.E.: He throws like a girl (but only when he's sad): Emotion affects sex-decoding of biological motion displays. Cognition 119(2), 265–280 (2011)
[17] Kirouac, G., Dore, F.Y.: Accuracy of the judgment of facial expression of emotions as a function of sex and level of education. Journal of Nonverbal Behavior 9(1), 3–7 (1985)
[18] Laban, R., Lawrence, F.: Effort. Macdonald and Evans (1947)
[19] Lang, P.J., Greenwald, M.K., Bradley, M.M., Hamm, A.O.: Looking at pictures: Affective, facial, visceral, and behavioral reactions. Psychophysiology 30(3), 261–273 (1993)
[20] Lewis, M.: Self-conscious emotions. American Scientist 83(1), 68–78 (1995)
[21] Lu, G., Shark, L.K., Hall, G., Zeshan, U.: Dynamic hand gesture tracking and recognition for real-time immersive virtual object manipulation. In: International Conference on CyberWorlds, pp. 29–35. IEEE (2009)
[22] McDonnell, R., Jörg, S., McHugh, J., Newell, F., O'Sullivan, C.: Evaluating the emotional content of human motions on real and virtual characters. In: Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization, pp. 67–74. ACM (2008)
[23] Measurand: Motion capture systems (2009), http://www.measurand.com
[24] Montagne, B., Kessels, R., Frigerio, E., De Haan, E., Perrett, D.: Sex differences in the perception of affective facial expressions: Do men really lack emotional sensitivity? Cognitive Processing 6(2), 136–141 (2005)
[25] Rotter, N.G., Rotter, G.S.: Sex differences in the encoding and decoding of negative facial emotions. Journal of Nonverbal Behavior 12(2), 139–148 (1988)

[26] Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39(6), 1161–1178 (1980)
[27] Samadani, A.: Questionnaire videos (2011), https://ece.uwaterloo.ca/~asamadan/JulyVideos.htm
[28] Samadani, A., Kubica, E., Gorbet, R., Kulić, D.: Perception and generation of affective hand movements. Submitted to International Journal of Social Robotics (2012)
[29] Samadani, A., DeHart, B.J., Robinson, K., Kulić, D., Kubica, E., Gorbet, R.: A study of human performance in recognizing expressive hand movements. In: 20th IEEE International Symposium on Robot and Human Interactive Communication, pp. 93–100 (2011)
[30] Spiegel, J., Machotka, P.: Messages of the body. Free Press, New York (1974)
[31] SPSS: Spss for windows, rel. 19.0. SPSS Inc., Chicago (2010)
[32] Stevens, J., Hamann, S.: Sex differences in brain activation to emotional stimuli: A meta-analysis of neuroimaging studies. Neuropsychologia 50(7), 1578–1593 (2012)
[33] Wagner, H.L., MacDonald, C.J., Manstead, A.S.: Communication of individual emotions by spontaneous facial expressions. Journal of Personality and Social Psychology 50(4), 737–743 (1986)
[34] Wallbott, H.G.: Bodily expression of emotion. European Journal of Social Psychology 28(6), 879–896 (1998)
[35] Wood, W., Rhodes, N., Whelan, M.: Sex differences in positive well-being: A consideration of emotional style and marital status. Psychological Bulletin 106(2), 249–264 (1989)

Vagueness and Dreams: Analysis of Body Signals in Vague Dream Telling

Laura Vincze, Isabella Poggi, and Francesca D'Errico

Department of Education,
Roma Tre University Rome, Italy
laura.vincze@gmail.com, {poggi,fderrico}@uniroma3.it

Abstract. The paper provides a conceptual definition of the notions of vagueness and approximation (a lack of detail or precision in the knowledge one has of something), hesitation and hastiness (the act of waiting before, or hurrying up while speaking), and overviews some reasons why people can be vague, approximate, hesitating or hasty. A study is presented in which participants tell a recent dream of theirs, and a qualitative analysis is proposed of the words, gestures and other bodily signals that communicate vagueness, approximation, hesitation, hastiness, and word search during dream-telling, by pointing out their semantic differences and the features that distinguish them.

Keywords: Vagueness, approximation, hesitation, hastiness, social signals, multimodal communication.

1 Introduction

When we talk to other people, whether arguing in a discussion or telling a story during small talk, we are bound to implicit norms of communication, like Grice's Cooperation Principle and the maxims of Quantity, Relation, Quality, Manner, that impose to provide sincere and unambiguous information, and not to tell not more nor less than what is relevant for our interlocutor. Sometimes, though, we do know the information we are providing is not totally certain, complete, accurate or detailed, and we acknowledge this by verbal, prosodic or gestural markers of uncertainty (Lakoff, 1973; Rowlands, 1995; Dral et al., 2011) or vagueness (Poggi & Vincze, 2012) that metacommunicate our "caveats" about the information conveyed. By doing so we in a sense apologize and justify ourselves for keeping below the threshold of minimum needed quantity and clarity of information.

Actually, two kinds of problems may hold in the information we are providing and the way we convey it: one concerning the cognitive properties of information itself, for instance its being vague or approximate, and one concerning our own cognitive activity while conveying it, like our hesitation or hastiness. This work deals with a set of signals that convey these two kinds of meaning: the vagueness or approximation of the information we are providing, and our own hesitation or hastiness in phrasing it. Investigating these signals enlightens the level of sophistication of our everyday

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 77–89, 2012.

communication but also provides insights about how sophisticated nuances of human communication might be sensed and simulated by Virtual Agents and Robots.

2 Vagueness and Connected Notions

The concept of vagueness has been so far investigated by language philosophers (Keefe, 2006) but while markers of uncertainty have been tackled no one has studied how people acknowledge their vagueness in discourse through multimodal communication. Recently, Poggi & Vincze (2012) provided a cognitive definition of vagueness as a property of the knowledge assumed about a certain topic: a lack of detail in what one knows about something. As we are vague, we do not have a detailed knowledge of the topic, but only general beliefs, and not ones on particular aspects of it.

Vagueness was contrasted to precision; the fact of having beliefs on each specific aspect of a topic, but distinguished from uncertainty, since we may have a vague knowledge, a vague idea, a vague remembering of something, but still be certain of it. Vagueness was distinguished from approximation, a lack of precision concerning quantitative aspects of the topic, as opposed to vagueness that concerns qualitative aspects of it: the former has to do with measuring, the latter with describing. Approximation is close to uncertainty and, like vagueness, opposed to precision, and precision is the opposite of both approximation and vagueness, but viewed from two different angles: quantity and quality.

Besides setting the conceptual differences between these phenomena, Poggi & Vincze (2012) looked at how we multimodally communicate meanings of vagueness and approximation during discourse, by defining "vagueness signals" as the verbal or bodily metadiscursive signals (Poggi, 2007) that convey the meaning "I am being vague". In general, metadiscursive signals reveal the Sender's goals concerning her/his discourse planning, i.e. what s/he considers important, what s/he affords to skip, and what logical links s/he states among parts of her/his plan. During discourse, if we want to convey we are being less detailed or accurate in some parts of it, e.g., because those parts are not so important in the economy of the whole discourse, we may do so by words, gestures, gaze or facial expressions. These are "vagueness signals", i.e. metadiscursive signals that convey "I deliberately choose to be vague about this".

The quoted work also singled out the characterizing features of signals for vagueness and approximation: "vagueness gestures" generally share the features of a basic and easy handshape (open hand, curve fingers, generally no protruded fingers), and curve movement trajectory. They are generally repeated, possibly in shape of a circle and with a cyclic form, and often involve movements of outward rotation (as opposed to the oscillation of approximation gestures), with low muscular tension and high fluidity. Moreover, they are sometimes accompanied by eyes looking upward or sideways, typical of someone who has not yet found the right concept, or by a grimace with lips lowered conveying "I don't know".

The motor features of curve handshape, outward rotation and low tension of the "metadiscursive vagueness gestures" metaphorically evoke the blurred and fuzzy

knowledge typical of vagueness, and can be seen as an embodied meaning of relaxation of someone who is aware of being vague, but feels s/he can afford to be so, and conveys: "I am relaxed since speaking of this is not so important, thus I can afford being vague".

"Approximation gestures" instead generally involve an oscillation of head and hands, with open stretched hands, sometimes with spread fingers.

Two main reasons were singled out why we may deliberately decide to be vague: ..: either we do not have detailed information about the topic (no power) or we are aware of details, but do not want to provide detailed information (no goal). In the former case, we lack information for being precise, in the latter, we deliberately decide not to provide precise information.

These two causes may combine in approximation: we may be approximate either because we lacks knowledge about the precise quantity at issue (no power), or because stating a different quantity than the actual one does not make a significant difference for either ourselves or our Interlocutor (no goal).

3 Cognitive Properties of Knowledge and Cognitive Processes in Communication

This work further investigates the signals of vagueness and approximation, while trying to go more in detail in the notions and expressive signals of three more phenomena: word search, hesitation, and hastiness.

We first assign these five areas to two different classes: vagueness and approximation are properties of the knowledge one conveys in communication, while word search, hesitation and hastiness are processes occurring when one has to transfer pieces of knowledge through communication.

The two further concepts of hesitation and hastiness are connected, as we shall see, to the notions of vagueness and approximation investigated so far. In a sense, hesitation is the opposite of hastiness, since as we hesitate we have the goal to take more time before doing something, while when we are hasty, we have the goal to hurry up not to lose time.

4 Hesitation

Hesitation can be defined as a non-action, or better, an action of waiting, of deliberately taking time – leaving time elapse – before doing something (something that in any case one has the goal, or is expected to do) either due to total lack of knowledge on what to do, or due to indecision between two or more actions. In fact one may hesitate because one utterly does not know what to do, but also because one has not yet made up one's mind on either pursuing a goal or not, or on which action to do to pursue some goal. If one offers me a pastry with cream I may be undecided on whether to eat the pastry or dieting; but if I already decided not to diet, I may be undecided on whether to choose a cream or a chocolate pastry .

Besides hesitating before doing some practical action, we can hesitate before some communicative action, like answering a question. In this case too, hesitation may be caused by indecision on whether to give the answer A or the answer B, by indecision on whether to answer or not, or finally by total lack of knowledge on what to answer.

In this work we focus on hesitation in communication, which can be defined as waiting before doing a communicative action. Let us make some hypotheses about the process that gives rise to phenomena of hesitation and hastiness, and to their communication. Hesitation in communication can be generated by at least three cases:

1. WORD SEARCH: I have the goal to convey some concept and search my mental lexicon for a right word to convey this; but I cannot retrieve the right word: this gives rise to a goal of waiting and taking time before uttering new words and generally going on in discourse
2. COMMUNICATIVE INDECISION: I have the goal to convey some concept, I search my mental lexicon, but I find more than one word: this gives rise to indecision, and consequently to a goal of making up my mind on which word to utter, which in its turn generates a goal of taking time
3. RETICENCE: At first I have the goal to communicate a particular content, or to find a particular word that conveys it; but then I evaluate the possible consequences of my communicating it, either in view of my own interests (selfish worry) or in view of my Interlocutor's interests (altruistic worry). This generates the goal of not uttering exactly that word, or phrasing that concept, but substituting it with a lie (in the selfish case) or, for example, with a euphemism (in the altruistic case). This necessity for further concept or word search triggers a need for taking time.

In the first two cases the goal of taking more time may be consciously and deliberately communicated, not only by the actual time elapsed before the next word, (i.e. longer than expected), but also by perceivable signals like filled pauses or so; in the third case the unexpected length of time elapsed may work as an informative signal (Poggi & D'Errico, 2012) – i.e., a signal that provides information to the receiver though not deliberately emitted by the sender. In both the selfish and the altruistic case, the Sender's goal is not to let the Interlocutor understand the reason for one's hesitation, nor, if possible, to make the hesitation itself observable and conspicuous.

Interestingly, vagueness may be a sub-case of case 1. Sometimes it is not that you really can't find a word, but you simply feel that word is not precise enough as you would like it to be; in this case you can hesitate just because you do not want to be vague.

5 Hastiness

We define hastiness as the goal of not losing time while doing something. One who is hasty is doing something, but in doing it, one does not want to bother to do things in a particularly precise or accurate manner. This is so in hasty communication too, where hastiness can be defined as the goal of not wasting time in the lengthy definition or description of contents of ongoing discourse.

Hastiness may be determined by some steady trait of the Speaker. As argued by Poggi (2007), personality can be seen as the fact of attributing more importance to some goals than to others. Therefore people's personality may determine their communicative style, i.e., their tendency to be more or less hasty, but also the respective communicative outputs of their hastiness. Let us describe various types of possible senders, the reasons and outcomes of their hastiness, and make some predictions about their hastiness signals.

4. VAGUE BY NATURE. Mr. X generally does not bother about being precise: the goal of being precise is not so important for X, so when competing with a possible goal of saving time in conversation, the latter wins: this triggers a goal of being contented with a low level of precision. Yet, X realizes he is keeping below the Grician norm of sufficient information, and metacommunicates his being brief and hasty.
5. PRECISE BY NATURE. Mr. Y, the opposite of Mr. X, generally likes being detailed: he attributes high importance to the goal of being precise. But in this case he cannot be precise, due to vagueness or imprecision in his very underlying knowledge: he then he communicates his imprecision, but since this is a blow to his own image of a precise person, he feels irritated. So we may predict a nuance of irritation leaking from the parameters of movement in his gestures.
6. SELFISH PRIVACY PROTECTOR. For Mr. Z, his goal of privacy is very important, while the goal of being precise is of medium importance. When he feels that going too much into detail might disrupt his goal of privacy, he decides to be imprecise; but instead of saying "well it's too private", he may mask his will of being imprecise by a will of being fast. He is in a sense forced to do so because, should he sincerely confess he does not want to go into details, this would anyway give hints to the interlocutor.
7. ALTRUISTIC EUPHEMISTIC PROTECTOR. Mr. K does like to be precise. Yet, when information can be painful or disrupting (Castelfranchi & Poggi, 1998), since he considers the goal of not hurting the other very important, he prefers to protect the Interlocutor by being imprecise. In this case too, like as for Mr. Z, to conceal that he is being imprecise only to avoid hurting, Mr. K pretends to simply be in a hurry: hastiness as a means for euphemism.

In the following we present a study aimed at analyzing verbal and multimodal signals of Vagueness, Approximation, Word Search, Hesitation and Hastiness.

6 Data Collection and Analysis

The analysis of body signals of vagueness, word search and approximation undertook by Poggi & Vincze (2012) was performed on a corpus of recorded oral examinations. Although during exams answering questions in a vague or blurred way definitely goes against one's interests, yet in this context students often do provide vague accounts of very precise issues. And while oral exams are rich in gestures, prevalently batons,

iconic and deictic, because in trying to be clear the students tend to employ every possible means (linguistic, kinetic and paralinguistic), body signals of vagueness are also present, though fewer than other types of signals.

To go deeper in the area of vagueness signals, we collected another type of corpus. Capturing instances in which people report about vague concepts required eliciting situations where one does not have precise knowledge/remembrance about the specific concepts at issue: such as in dream-telling. The events in a dream are, by their nature, often confuse and vague, as one does not only mismatch entities (people, places) with one another by attaching some attributes of *x* to *y*; but sometimes, one does not have precise remembrance about some relevant attributes of these entities. Therefore we assumed that people reporting their dreams might employ fuzzy gestures with imprecise trajectories.

We designed an experiment where 25 students in Education Sciences, 23 females and 2 males, between 20 and 30, were asked to tell a recent dream. While telling their dream to an interviewer, the participants' body movements were recorded by a digital Panasonic camcorder. Participants were seated on a chair without armrest and only their upper body (trunk, arms, hands and head) were video-recorded. When participants, during their narration, mentioned vague remembrance or vague knowledge of events happening in the dream, the interviewer would ask them to detail those aspects of the dream with the intent of "raising" (Gianturco, 2004) the performance of (possibly vague) gestures accompanying vague memories.

A total of 25 fragments of dream telling were collected, consisting of approximately 5 minutes each. The verbal behaviour of participants was transcribed by taking into account the intonation unit (IU) as the basic unit of transcription. The intonation unit is a prosodic unit in natural discourse, a speech segment that falls into a single coherent intonation contour, and is sometimes separated by pauses at the beginning and the end (Chafe 1987, Du Bois et al. 1992).Transcribing the data in IUs, each IU lined up on a separate line, helps readers to more easily grasp the pauses in speech.

Before this transcription, the vide-recorded data were first viewed on mute mode to avoid bias from the verbal context. When items of gestures or facial expressions possibly conveying vagueness meanings were singled out, the video fragment was reviewed on voiced mode, transcribed and later coded by two independent coders. All the body signals conveying vagueness, approximation, word search, hesitation and hastiness were transcribed and analyzed in an annotation scheme of multimodal communication (of the type of Poggi, 2007). For each signal we annotated: 1. concomitant verbal behaviour, 2. analysis of the signal (for a gesture, its handshape, place, orientation, and the parameters of movement, such as direction, path, tension, amplitude, fluidity, repetition); 3. possible concomitant body behaviour, like gaze, smile, posture; 4. the meaning attributed to the signal at hand. Based on such annotation, each signal was coded as one of vagueness, approximation, word search, hesitation or hastiness, and a hypothesis was made as to the reason (no-goal or no-knowledge) for the participant to be vague, approximate or other in that context.

7 A Qualitative Analysis of Signals of Vagueness, Approximation, Word Search, Hesitation and Hastiness

Let us illustrate some cases extracted from the dream telling corpus.

In the first video a female student tells us about her attempts to rescue a crow, in her dream of the previous night. She often stops and takes time to search for the right words, accompanying these moments by word search gestures, or else she interrupts the sentence flow because of vague remembrance, often making with loose, rotating gestures. Let us first analyze and distinguish between instances of word-search, approximation and hastiness.

7.1 Word-Search Gestures

When searching the most appropriate term for the particular context at hand, the speaker tends to fill her pause in speech either with fillers such as *come dire* "how to say" or by prolonging the preceding vowel. Jerky gestures, sometimes with rotating wrist or oscillating stretched fingers, often occur, symbolically representing the speaker's search for the right term.

(1)
00.40 *E mi ricordo appunto che mentreeeee […]*
Questo sogno
è particolare perché il giorno dopo mentre uscivo di casa
ho visto
un corvo di fronte aaaaaaal
roseto
di casa mia […]
*(*And I remember that whiiiiiiile […] This dream is special because the day after while Iwas going out of myyyyy house I saw a crow in front ooooof the rosebush in my garden[…]).

While in the first case of prolonged vowel (*mentreeeee*= whiiiile) the Speaker abandons the sentence and the word search (or memory search) and starts all over again by telling about the dream in general "*Questo sogno è particolare*" (This dream is special), in the second case (*aaaaaaal*, = ooooof) the Speaker successfully concludes her search of the word *roseto* (rosebush) and iconically represents it.

It might be that every time one is searching a word one prolongs the previous vowel, with the only difference being that in the latter case the search was successful, while in the former the speaker decides to go on with the narrative, even if the search of that word has not been successful.

7.2 Approximation Gestures

To detect fragments where the speaker was being approximate we often relied on the "lexical affiliates" (Hadar & Butterworth 1997) of approximation gestures: typically, adverbs communicating approximation, such as "more or less", "about", "around", "almost" "a bit more than", "approximately", "roughly". To communicate approx-

imate quantities by body signals, our participants perform oscillating gestures where the hand goes from right to left as windscreen wipers, but without the wrist rotation typical for vagueness gestures. The hand oscillation from right to left and back is symbolically parallel to the concept of "around". It is as if the speaker sets out a point in space where the precise quantity is, and by oscillating around that point, she communicates the impossibility of reaching that very point. This impossibility of communicating a precise quantity stems either from a no-knowledge or from a no-goal cause (she may either ignore the precise quantity or not consider it important for the goals of the discourse). In fact, the very meaning "more or less" testifies that a higher or lower amount does not really make a difference. In fact, as already mentioned, approximation is a lack of precision concerning quantitative aspects of the topic.

Let us see a gesture of approximation (or, strictly speaking, of unspecificity[1]) that appears in concomitance with lexical affiliates like *una specie di* (a sort of) and *come* (like).

(2)
Però era tipo una specie di discoteca c'era la musica fortissima
(But it was a sort of disco there was very loud music)

Tipo una specie di discoteca (a sort of (disco)) implies that the concerned place does not fulfill all the conditions to be considered a disco *comme il faut*. "A sort of" communicates that the place at issue cannot reach the threshold of defining features according to which one can call a disco a disco: it is less than that. How much less, one cannot tell, one can only be approximate, as the girl in the video. She is telling her dream about being in a (sort of) disco and while stating *una specie* (a sort), her left hand goes from right to left as windscreen wipers, without any wrist rotation.

In another example a student tells the interviewer about her dream about the end of the world. When describing the color of the sky as she perceived it in the dream, she uses several hedges to warn the listener of her incapacity of being more specific. A first hedge is *colore tipo* (a colour like/kind of); another one is a neologism: *un cielo giallastro arancionato* (a yellowish oranged sky). In Italian *–astro* is a consolidated

[1] We tend to distinguish lack of precision due to uncertainty concerning quantities or intensities (the case of approximation) from lack of precision due to lack of detailed knowledge of the qualitative aspects of the topic. This latter case, depending on whether the lack of detailed knowledge stands in the field of definition or in one of description, results in either unspecificity or vagueness, respectively. This subtle issue will be investigated in depth in a later work; for now we only mention that the three last examples belong, in our view, to the category unspecificity, and not so much approximation, since the impossibility of the Speaker is one of *defining* the object at issue due to lack of knowledge on its qualities.

Some hedges (Lakoff, 1973), the words to acknowledge lack of precision, stand at the border between unspecificity and approximation, as they can be used both for quantities/intensities and for qualities. "Almost" (it. *quasi*) is one of them: one can both say: "There were almost 20 people at the party" (where "almost" clearly has a quantitative valence) and "He's almost a man" (where "almost" refers to qualitative aspects, i.e. getting qualitatively close to being a man). As it often happens, the reign of designation gets extended from physical, concrete objects to abstract ones as well; this is also the case for "almost".

suffix for colour derivates, that when added to the end of a color name communicates a close but not total similarity to that colour (*giallo-giallastro* = yellow-yellowish; verde-verdastro = green-greenish): it communicates closeness to the original color but at the same time a difference from. Unlike *–astro*, the suffix *-ato* is not morphologically productive as meaning "something like". In our case, then, to tell the color of a sky that she cannot describe precisely, our participant creates a neologism, a new word with a suffix added to a root to which it is not usually attached: she adds the non-productive suffix (*-ato*) to a root to which it is not usually attached. And using a completely new term (*arancionato*) is well adapted to the necessity of conveying a somewhat unknown content, like the strange things we dream in our dreams. The speaker herself, not completely satisfied with her lack of specificity and also insecure whether her neologism created on the spot will be accepted by the interlocutor, turns her nose up in a grimace and lowers her head and shoulders. To turn nose up, a part of the expression of disgust, is often used to express a negative evaluation; so she is herself evaluating her own neologism badly; while lowering head and shoulders is a submissive posture by which she is apologizing for it.

Approximation can also be a subgoal of a further goal like not wasting time for irrelevant explanations. In communication, when the interlocutor asks a clarification question, he expects an answer containing the due quantity of explanation. The speaker may then have two, possibly incompatible, goals: one of providing due clarifications and one of not wasting time in irrelevant details. This clash between the two goals can lead to the goal of being imprecise to save time, consequently causing a lack of precision in the speaker's answer (which becomes either approximate or vague). Both approximation and vagueness can therefore also stem from a goal of not wasting time, not only from uncertainty or lack of knowledge.

We now see how approximation and vagueness can be connected to the concept of hastiness. As seen in section 5, hastiness stems from one's not wanting to waste time. In our next example approximation combines with hastiness. The girl of the "crow" dream is now talking about the people who were present in her dream while she was trying to help the crow. In the dream there was also a boy that she knew.

(3)
e io questo ragazzo conosco,
èèèèèèè un
diciamo un amico mio
(and I know this boy he iiiiiiiis a let's say a friend of mine)

The hedge *diciamo* (let's say) works, in this case, as an alert signal warning of something that we might call "conceptual approximation": the Speaker's statement should be taken with a grain of salt, that is, the interlocutor should maintain a degree of skepticism about its truth. Saying that someone "is, let's say, a friend" signals that from a quantitative point of view that person does not have all the prototypical qualities to deserve being called a friend. Nonetheless, the Speaker does not wish to insist on this matter because irrelevant for the goals of the present discourse. Hence, the meaning of that statement may be paraphrased as: "It is not really so but it does not matter. Let's not get into that".

In our case, before saying "let's say, a friend of mine", the participant hesitates and takes time to decide how to call the boy at issue. Her hesitation is marked by a *pause* in speech flow, a pause filled by the *prolongation of the monosyllabic verb* "*è"* (is). The speaker decides not to waste time to get into detail on the degree of friendship between herself and the boy, and she accompanies her words with *rapid rotating movements*. Velocity implies tension, so we are not dealing with a vagueness gesture here. It is a metaphorical gesture designating big quantities which are accepted on the basis of no special control: something like saying "grossly speaking". In fact, this is where the inattention to details comes from. As you are worried about wasting time, details are not your main concern.

7.3 Hastiness Gestures

In the previous example the participant gives approximate information not to waste time in irrelevant details. Later, in the same video, she gives an example of hastiness due to slight irritation: as the interviewer insists on a more detailed description of the boy at issue (how he looked like), she mentions a quite distinctive feature of the boy (he had a pony-tail), but tries to make this detail pass as non relevant, one not worth wasting time talking about it. She also looks somewhat irritated since she performs a *tense and rapid gesture with both hands open, facing upwards and intersecting each other moving as scissors*. In this gesture, the *upward orientation of palms* evokes the visual metaphor of showing one's bare hands, paraphrasable both as "I'm sincere, I do not hide anything from you" and as "What I show you (bare hands) is all I have". These two possible interpretations may be combined in the meaning of "I do not know anything more than this" while the movement of *both hands intersecting with each other* like two blades of scissors conveys "let's cut the unimportant details". Thus, globally the gesture can be interpreted as a cutting short strategy. The student's *head comes forward* as if asking "What do you want me to say?". Her hastiness might derive either from not wanting to pause over such personal things (see the case above "selfish privacy protector") or from having imprecise or vague knowledge ("precise by nature").

7.4 Vagueness Gestures

The "crow" dream gives us two examples of the two possible reasons for being vague: no knowledge and no goal. Let us see the former: a gesture motivated by no knowledge, not only of a word but even of the concept.

While word search instances are, in most cases, resolved by finding the wanted word (consequently prosodically and/or gesturally emphasized), vague concepts are difficult to seize and most speakers give up in the middle of the description, abandoning the task and leaving the sentence suspended. In one fragment the girl unsuccessfully tries to shed light on aspects of her dream. After saying *mi ricordo appunto che mi ricordo* (I in fact remember that I remember), she *makes a pause* and performs a vagueness gesture with *both hands open, right hand performing circular and loose*

rotating movements: remembrance is vague. She therefore gives up remembering and moves forward with the dream account.

In another example, vagueness is a conscious choice and comes after a moment of hesitation about whether or how to mention certain unpleasant issues. This fragment comes immediately after she mentioned the crow in the rosebush:

(4)
[...] che stava praticamente ha preso un uccellino
e se l'èèèèè praticamente portato via
(which was standing it virtually caught a little bird and virtually took it away)

The Speaker uses a vague euphemism to convey the idea, as if preferring not to mention certain macabre things such as the crow killing the little bird. While prolonging the vowel *èèèè* she *rotates right hand open with rapid and tense movements*. The Speaker performs a light embarrassment *smile* while *fixedly gazing at the interlocutor*. We may notice that a conscious choice of being vague is not characterized by *loose movements* as in the previous case where the Speaker dealt with vague remembrance, but by *rapid and tensed rotations*. In fact, when deciding whether to tackle unpleasant information, the Speaker makes a conscious choice which eventually may lead not to mention certain things (no goal). Decision making can sometimes be a tense and stressing process, not at all loose and relaxing as when we give up search and admit – to oneself and others – that we do not have that specific piece of information.

8 Concluding Remarks

From a conceptual point of view, the five notions tackled in our work (vagueness, approximation, hastiness, hesitation) are all connected to one another. Three of them share an element of lack of precision (approximation and vagueness stand for quantitative and qualitative lack of precision, respectively, while hastiness stems from the goal of not wasting time by being too precise). Hesitation only (here we refer to hesitation in word search) does not stem from a lack of precision, but in a way, from a desire of precision, and in this sense it is the opposite of hastiness.

As we have seen, word search may determine hesitation in discourse flow. On the contrary, hastiness, being characterized by a lack of interest in doing things in a particularly precise or accurate manner, may trigger approximation. One who is required to give some information but is in a hurry and does not want to waste time, may settle with providing an approximate information and signal this by an approximation gesture. In our corpus, we singled out gestures belonging to all these categories. Every gesture having a fuzzy-round, oscillatory, rotating or jerky trajectory was taken into consideration and analyzed. Four categories of gestures were singled out as belonging to this trajectory description, namely: vagueness (fuzzy-round), approximation (oscillatory), word search (rotating) and hastiness gestures (jerky).

Some categories of gestures share common parameters: vagueness and word search have the rotating movement in common, but while in vagueness the rotation is loose,

in word-search the Speaker performs more rapid rotations (possibly due to irritation for not finding the right word and/or to time constraints in conversational turn-taking). Other times, within the same gestures category there may be a difference in gesture parameters. This is the case of vagueness gestures, that may either be *loose* and *slow* with *averted gaze* (*eyes up* in the sky or *lowered* often with *tight eyelids* possibly conveying effort in focusing) or rapid and tensed, accompanied by *direct eye contact* and sometimes by *smile*. The former is the case of vagueness gestures performed when there is no remembering (no knowledge), while the latter is the case of vagueness gestures performed when the Speaker prefers to remain vague and allusive in the expression of his communicative content. The former may occur either in absence of vocalizations or, if vocalizations are present, the gesture is accompanied by a prolonged *mmmmmmm* sound (it signals the ongoing cognitive process while trying to shed light on the vague remembrance). *Sentences* are sometimes left *suspended* as the Speaker does not know how to continue the description (no knowledge). In the latter (no goal) case, the produced vocalization is the prolongation of the last preceding vowel or of a [ə]. This particular case is very similar to hesitation. In fact while hesitating one mentally calculates the advantages and disadvantages – for oneself, interlocutor or third person – of speaking or not speaking and while reasoning on this, one fills in this hesitation moment by a *prolonged [ə]*.

This paper stems from the desire of investigating the concept of vagueness in body communication and it represents a further step towards the understanding of the so-called vagueness gestures and the speakers' goals in performing them. It is nonetheless a preliminary research, tackling only qualitative aspects of vagueness communication. In a future study we will investigate quantitative differences between vagueness, approximation, word-search and hastiness gestures in the two corpora of dream-telling and oral examination, and possibly in other corpora.

Acknowledgements. Research supported by SSPNet Seventh Framework Program, European Network of Excellence SSPNet (Social Signal Processing Network), Grant Agreement N.231287.

References

1. Lakoff, G.: Hedges: A study in meaning criteria and the logic of fuzzy concepts. Journal of Philosophical Logic 2(4), 458–508 (1973)
2. Chafe, W.: Prosodic and functional units of language. In: Edwards, J.A., Lampert, M.D. (eds.) Talking data: Transcription and coding methods for language research. Lawrence Erlbaum Associates, Hillsdale (1992)
3. Du Bois, J.W., Schuetze-Coburn, S., Paolino, D., Cumming, S.: Discourse transcription. In: Thompson, S.A. (ed.) Santa Barbara Papers in Linguistics, vol. IV. UCSB (1992)
4. Dral, J., Heylen, D., den Akker, R.: Detecting Uncertainty in Spoken Dialogues: An explorative research to the automatic detection of a speakers' uncertainty by using prosodic markers. In: Ahmad, K. (ed.) Affective Computing and Sentiment Analysis, Text, Speech and Language Technology, vol. 45, pp. 67–77. Springer, London (2011)
5. Rowland, T.: Hedges in mathematics talk: Linguistic pointers to uncertainty. Educational Studies in Mathematics 29, 327–353 (1995)

6. Keefe, R.: Theories of Vagueness. Cambridge University Press, Cambridge (2000)
7. Castelfranchi, C., Poggi, I.: Bugie, finzioni, sotterfugi. Per una scienza dell'inganno. Carocci Editore, Roma (1998)
8. Gianturco, G.: L'intervista qualitativa. Guerini Editore, Torino (2004)
9. Hadar, U., Butterworth, B.: Iconic gesture, imagery and word retrieval in speech. Semiotica 115, 147–172 (1997)
10. Poggi, I.: Mind, Hands, Face and Body. A goal and belief view of multimodal communication. Weidler Buchverlag, Berlin (2007)
11. Poggi, I.: Mind markers. In: Rector, M., Poggi, I., Trigo, N. (eds.) Gestures. Meaning and use, pp. 119–132. University Fernando Pessoa Press, Oporto (2002)
12. Poggi, I., D'Errico, F.: Social signals. A framework in terms of goals and beliefs. In: Poggi, I., D'Errico, F., Vinciarelli, A. (eds.) Foundation of Social Signals. From Theory to Application. Special Issue of CognitiveProcessing. Springer, Heidelberg (2012)
13. Poggi, I., Vincze, L.: Communicating vagueness by hands and face. In: Proceedings of the ICMI Workshop on Multimodal Corpora for Machine Learning, Alicante (2012)

Computing and Evaluating the Body Laughter Index

Maurizio Mancini, Giovanna Varni, Donald Glowinski, and Gualtiero Volpe

InfoMus - University of Genoa
{maurizio.mancini,donald.glowinski}@dist.unige.it,
{giovanna,gualtiero}@infomus.org

Abstract. The EU-ICT FET Project ILHAIRE is aimed at endowing machines with automated detection, analysis, and synthesis of laughter. This paper describes the Body Laughter Index (BLI) for automated detection of laughter starting from the analysis of body movement captured by a video source. The BLI algorithm is described, and the index is computed on a corpus of videos. The assessment of the algorithm by means of subject's rating is also presented. Results show that BLI can successfully distinguish between different videos of laughter, even if improvements are needed with respect to perception of subjects, multimodal fusion, cultural aspects, and generalization to a broad range of social contexts.

1 Introduction

Traditional Human Computer interfaces are frequently perceived as "cold, incompetent, and socially inept". According to Zeng and colleagues, this results from the fact that they ignore the user's affective state and consequently miss a key component of human-human communication [1]. This is why, in the last years, progress was made toward the creation of emotional Human-Computer interfaces, see for example Affective Computing [2] and Kansei Information processing [3].

Laughter is a relevant component in human-human nonverbal communication and it is a powerful trigger for facilitating social interaction. Indeed, Grammer [4] suggests that it conveys signals of social interest and reduces the sense of threat in a group [5]. Further, laughter seems to improve learning of new activities from other people [6] and facilitates sociability and cooperation [7].

For the above reasons, the newly started EU-ICT FET Project ILHAIRE (http://www.ilhaire.eu) aims to investigate how machines can decode laughter (i.e., to know when the user is laughing, to measure intensity of laughter, to distinguish between different types of laughter) and also how Embodied Conversational Agents can communicate laughter.

In our work, we mainly focus on the detection and on the analysis of the movement descriptors (e.g., speed, direction, periodicity, and so on) that are deemed to characterize laughter. Very few researchers investigated the role that body plays in human laughter, even if all of them agree that body configuration and dynamics contribute to the communication of different types of laughter.

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 90–98, 2012.

Ruch and Ekman [8] observed that laughter is often accompanied by one or more (i.e., occurring at the same time) of the following body behaviors: "rhythmic patters (five pulses per second)", "initial forced exhalation", "rock violently sideways, or more often back and forth", "nervous tremor ... over the body", "twitch or tremble convulsively". Becker-Asano and colleagues [9] observed that laughing users "moved their heads backward to the left and lifted their arms resembling an open-hand gesture". De Graaf [10] observed that laughing consists of a deep inspiration followed by a rapid convulsive expiration whereas de Melo et al. [11] implemented a virtual agent that "convulses the chest with each chuckle". Finally, Markaki and colleagues [12] analyzed laughter in professional meetings: the user laughs "accompanying the joke's escalation in an embodied manner, moving her torso and laughing with her mouth wide open" and "even throwing her head back".

A pioneering system including automatic detection of laughter is the *Affective Multimodal Mirror* [13] [14]. This system "tries to induce positive emotions in users by showing a distorted ("funny") representation of their face" [13]. The system senses and elicit laughter, based on a vocal and a facial affect-sensing module, whose outputs are integrated by a fusion module.

The above studies suggest that it should be possible to develop systems for automatic detection of laughter and even differentiate between different types of laughter (e.g., hilarious vs. social [12]). In this paper, we present a preliminary work in this direction in the framework of the ILHAIRE Project: we conceived and implemented the Body Laughter Index (BLI), an index that, by combining together movement descriptors, allows to automatically determine whether a user is laughing or not. We also describe a pilot evaluation study we conducted on the BLI.

2 Computation of the Body Laughter Index

Figure 1 depicts the software architecture for computing BLI. Next subsections provide details on the major software modules. All of them have been implemented in the EyesWeb XMI platform (`http://www.eyesweb.org`) and in the EyesWeb Expressive Gesture Processing Library [15]. EyesWeb is a software platform that allows developers to implement software modules for automatic analysis of user's expressive movement in an intuitive, visual way.

Based on the above literature, since laughter implies deep breathing (e.g., [10]) and possible rhythmic patterns, the initial set of descriptors taken into account for developing BLI includes shoulders correlation and energy of body movement, integrated with a measure of periodicity of movement.

2.1 Tracking of Head and Shoulders

Starting from an input video source (e.g., recorded video or camera), we detect and track the 2D position of user's head and shoulders. Head and arms movements are useful hints to detect one's affect [16]. We manually identify the

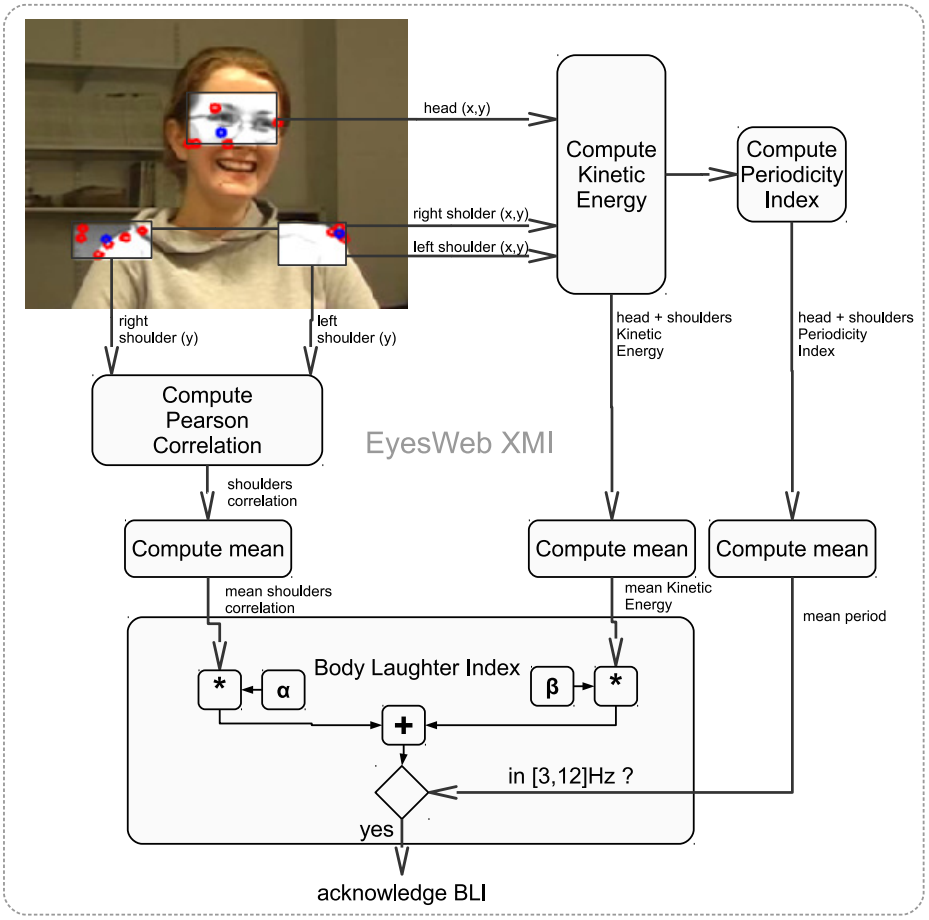


Fig. 1. The software architecture for computing BLI. Firstly, tracking of head and shoulders is carried out: the cloud of red points determines the Regions Of Interest (ROIs) head and shoulders are located in. The blue dots are the geometrical barycenters of each cloud. The boxes are the major software modules extracting and processing movements descriptors.

Regions Of Interest (ROIs) user's head and shoulders are located in (see the light areas in Figure 1). Standard computer vision tracking techniques are applied to each ROI, resulting in a cloud of points for each of them (see the red dots in Figure 1). Then we compute the geometrical barycenter of the cloud (see the blue dots in Figure 1) and we extract its x and y coordinates.

2.2 Low-Level Descriptors

We extract two low-level descriptors of the head and shoulders movement: *kinetic energy* and *correlation of shoulders movement*.

– *Kinetic energy* (E) is computed from the speed of the head (v_h), of each shoulder's barycenter (v_{ls} and v_{rs}), and their percentage masses (m_h, m_{ls}, and m_{rs}). These are derived from anthropometric tables as referred by [17]. In particular, kinetic energy is computed as:

$$E = \frac{1}{2}\sum_{i=1}^{3} m_i v_i^2 = \frac{1}{2}(m_h v_h^2 + m_{ls} v_{ls}^2 + m_{rs} v_{rs}^2) \quad (1)$$

– *Correlation of shoulders' movement* (ρ_s) is computed as the Pearson correlation coefficient between the vertical position of the user's left shoulder and the vertical position of the user's right shoulder. Vertical positions are approximated by the y-coordinate of each shoulder's barycenter extracted as mentioned above.

2.3 Periodicity Index

Kinetic energy is serialized in a sliding window time-series having a fixed length. *Periodicity Index* is then computed on such time-series. The Periodicity Index (PI) is a real-time implementation of the Periodicity Transforms described in [18]. The input data is decomposed into a sum of its periodic components by projecting data onto periodic subspaces. Periodicity Transforms also provide a measure of the relative contribution of each periodic signal to the original one. Among many algorithms for computing Peridiocity Transforms, we choose *mbest*. It determines the m periodic components that, subtracted from the original signal, minimize residual energy. With respect to the other algorithms, *mbest* also provides a better accuracy and does not need the definition of a threshold. Figure 2 shows an example of computation of the Periodicity Index in EyesWeb for the following input signal:

$$I(t) = sin(t) + N(t) \quad (2)$$

where $N(t)$ is a uniform random function generating values in $[0, 0.6]$ to simulate random noise. Such a range for $N(t)$ is chosen so that the noise is strong enough for simulation, but not so strong to destroy the original signal. The Periodicity Index value for such an input function is 20 frames, as Figure 2 shows.

2.4 Body Laughter Index

As mentioned above, the *Body Laughter Index* (BLI) stems from the combination of the averages of the low-level descriptors, integrated with the Periodicity Index. Such averages are computed over a fixed range of frames. However such a range could be automatically determined by applying a motion segmentation algorithm on the video source. A weighted sum of the mean correlation of shoulders' movement and of the mean kinetic energy is carried out as follows:

$$BLI = \alpha\overline{\rho_s} + \beta\overline{E} \quad (3)$$

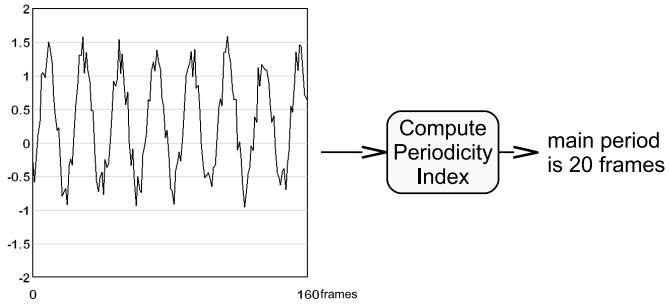


Fig. 2. An example of Periodicity Index computation: the input time-series (on the left) has a periodicity of 20 frames

As reported in [8], rhythmical patterns produced during laughter usually have frequencies around $5Hz$. In order to take into account such rhythmical patterns, the Periodicity Index is used. In particular, the computed BLI value is acknowledged only if the mean Periodicity Index belongs to the arbitrary range $[\frac{fps}{8}, \frac{fps}{2}]$, where fps is the input video frame rate (number of frames per second).

2.5 Example

We ran our algorithm for BLI computation on 8 short input videos at 25 fps taken from: (1) a previously recorded video corpus named "The Belfast Induced Natural Emotion Database", collected by the Queen's University of Belfast [19]; (2) the YouTube website (videos generated with the Skype Laughter Chain application, `www.skypelaughterchain.com`). The videos show users laughing while watching funny images on TV. They smile and laugh, tilting their head and producing rhythmic body movements.

Table 1 summarizes the results: the first column reports the video *id*; the second and third columns show the average values of the low-level descriptors (mean kinetic energy and mean Pearson correlation of shoulders' movements); the fourth column shows the computed BLI value; the last column reports the mean Periodicity Index.

In this example, parameters for BLI were set to $\alpha = 0.7$ and $\beta = 0.3$, respectively. These are arbitrary values, argued from the literature reported in Section 1. An in-depth study for optimal values of these parameters will be needed in future work.

3 Evaluation of the Body Laughter Index

BLI was also tested on 8 participants that watched and rated the 8 videos stimuli of Section 2.5. The stimuli were randomized using a balanced latin square. Participants were asked to rate on the following two 3-point Likert items: **Q1** "Did you perceive energetic body movements, involving shoulders rhythmically moving together?" and **Q2** "How fast was the rhythmic movement you perceived?".

Table 1. An example of computation of BLI. Kinetic energy $\overline{E}$ ranges in $[0, +\infty)$, correlation of shoulders' movements $\overline{\rho_s}$ ranges in $[-1, 1]$, BLI ranges in $[0, +\infty)$, and Periodicity Index $\overline{PI}$ ranges in $[0, w]$, where w is the time window length in frames.

Video id	$\overline{E}$	$\overline{\rho_s}$	BLI	$\overline{PI}$
1	40.7472	0.312	12.44256	16.2778
2	172.6268	0.358	52.03864	16.5362
3	117.4252	0.3508	35.47312	19.6532
4	14.458	0.6982	4.82614	7.874
5	0.5064	0.3092	0.36836	10.7522
6	0.1112	0.0664	0.07984	6.8234
7	250.8674	-0.2226	75.1044	18.9312
8	2.1034	0.5064	0.9855	10.293

These items were aimed at an initial assessment of BLI and of its components. Both items were rated from 0 (*not at all*) to 2 (*very much/fast*).

3.1 Video Samples

We first checked whether the ratings between videos are significantly different from one another. In other terms, we aimed at checking whether the 8 video samples, submitted to the participants, offered a sufficiently variable level of laughs for the pilot. We ran a Friedman test to observe possible differences between the participants' ratings for items Q1 and Q2 (see Figure 3).

Results show a significant effect for item Q1, $\chi^2(7, n = 8) = 16.492$, $w = 1.4$, $p < .05$, but no effect for item Q2, $\chi^2(7, n = 8) = 12.388$, $w = 1.24$, $p > .05$ ($p = .089$).

Post-Hoc tests were conducted to put in evidence possible differences between videos with respect to their Q1 ratings. The Bonferroni correction was applied to the levels of statistical significance (p-values) to control the inflation of type 1 error probability due to multiple comparisons. A significant difference was found between the ratings of video 2 and video 3 ($p = .032$).

3.2 Correlation of Movement Descriptors with Participants' Ratings

We were interested in evaluating Periodicity Index and Body Laughter Index with respect to the participants' ratings. We conducted a set of bivariate Kendall's tau-b correlations, whose results are shown in Table II. Findings show

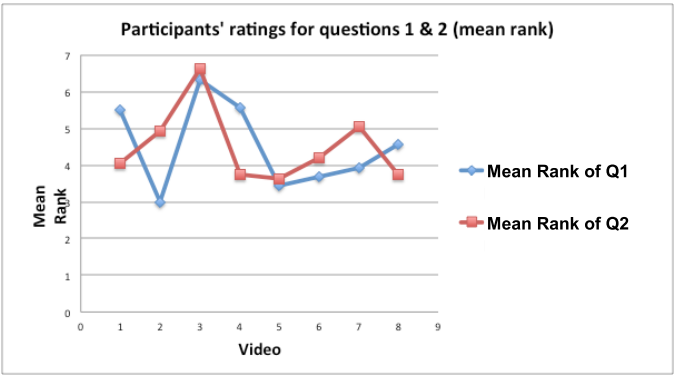


Fig. 3. Participants' ratings for Q1 and Q2

Table 2. Computation of bivariate Kendall's tau-b correlation between movement descriptors and participants' ratings

		Participants' rating	
		Q1	Q2
Descriptors	BLI	.07	-.09
	PI	-.14	-.25

the highest negative relationship between the Periodicity Index (PI) and Q2: $\tau = -.25$. Smaller relationships were also found between PI and Q1, $\tau = .14$, and between Body Laughter Index (BLI) and Q1, $\tau = .07$, and Q2, $\tau = -.09$.

4 Conclusion

In this paper we presented the implementation and evaluation of the Body Laughter Index, a body descriptor of laughter. Evaluation results show that some improvements are needed to reach successful automatic detection of laughter. The outcomes of BLI computation, reported in Table 1, indicate that BLI, combined with the Periodicity Index, allows us to successfully distinguish between different videos of laughter. However the evaluation of these videos by human participants, reported in Table 2, reveals that BLI and PI only partially match human perception. A possible reason is that laughter is a complex construct depending upon many features, as demonstrated by several studies.

In the future, in the framework of the EU-ICT FET Project ILHAIRE, we aim to carry out multimodal (audio, face, and body) fusion of descriptors: if audio signals analysis, facial expression detection and BLI computation agree with a high statistical significance, then we could claim that the user is laughing. We also aim to automatically differentiate between, for example, hilarious and

social laughter. Moreover, cultural aspects need to be considered as modulators of the interpretation of human movement. The resulting multimodal fusion will be assessed with a new set of experiments and the concerning evaluation.

An important issue to be taken into account is the context (activity which is performed, personality of the user, social environment) for laughter detection. Whereas BLI was computed with reference to a specific context (watching funny images on TV) and was evaluated in laboratory conditions, more research is needed to assess to what extent it is able to generalize to other, more general contexts.

From the implementation point of view, we aim to detect user movement with a higher resolution and more reliable systems, enabling to distinguish between different body parts (head, shoulders, and so on), such as Qualisys Mocap (http://www.qualisys.com) and Microsoft Kinect (http://www.xbox.com). An initial real-time implementation of BLI from live video input, using color tracking techniques, was developed and tested at the eNTERFACE'12 Summer Workshop on Multimodal Interfaces (Metz, France, July 2012).

Acknowledgments. We would like to acknowledge I. Sneddon and G. McKeown from Queen's University of Belfast for allowing us to use videos from the Belfast Induced Natural Emotion Database. This work was partially supported by the FP7 EU-ICT FET Project ILHAIRE (Incorporating Laughter into Human Avatar Interactions: Research and Experiments, http://www.ilhaire.eu).

References

1. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1), 39–58 (2009)
2. Picard, R.: Affective Computing. MIT Press (1997)
3. Hashimoto, S.: Kansei as the third target of information processing and related topics in japan. In: Proceedings of the International Workshop on KANSEI: The Technology of Emotion, pp. 101–104 (1997)
4. Grammer, K.: Strangers meet: Laughter and nonverbal signs of interest in opposite-sex encounters. Journal of Nonverbal Behavior 14(4), 209–236 (1990)
5. Owren, M.J., Bachorowski, J.-A.: Reconsidering the evolution of nonlinguistic communication: The case of laughter. Journal of Nonverbal Behavior 27, 183–200 (2003)
6. Fredrickson, B., et al.: The broaden-and-build theory of positive emotions. Philosophical Transactions - Royal Society of London Series B, 1367–1378 (2004)
7. Dunbar, R.I.M.: Mind the gap: Or why humans are not just great apes. In: Proceedings of the British Academy, vol. 154 (2008)
8. Ruch, W., Ekman, P.: The expressive pattern of laughter. Emotion, Qualia, and Consciousness, 426–443 (2001)
9. Becker-Asano, C., Kanda, T., Ishi, C., Ishiguro, H.: How about laughter? perceived naturalness of two laughing humanoid robots. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, pp. 1–6. IEEE (2009)
10. De Graaf, V.: Human Anatomy, 6th edn. McGraw-Hill, New York (2002)

11. de Melo, C.M., Kenny, P., Gratch, J.: Real-time expression of affect through respiration. Computer Animation and Virtual Worlds 21(3-4), 225–234 (2010)
12. Markaki, V., Merlino, S., Mondada, L., Oloff, F.: Laughter in professional meetings: the organization of an emergent ethnic joke. Journal of Pragmatics 42(6), 1526–1542 (2010)
13. Shahid, S., Krahmer, E., Swerts, M., Melder, W.A., Neerincx, M.A.: You make me happy: Using an adaptive affective interface to investigate the effect of social presence on positive emotion induction. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, pp. 1–6 (September 2009)
14. Melder, W.A., Truong, K.P., Uyl, M.D., Van Leeuwen, D.A., Neerincx, M.A., Loos, L.R., Plum, B.S.: Affective multimodal mirror: sensing and eliciting laughter. In: Proceedings of the International Workshop on Human-centered Multimedia, HCM 2007, pp. 31–40. ACM, New York (2007)
15. Camurri, A., Mazzarino, B., Volpe, G.: Analysis of Expressive Gesture: The EyesWeb Expressive Gesture Processing Library. In: Camurri, A., Volpe, G. (eds.) GW 2003. LNCS (LNAI), vol. 2915, pp. 460–467. Springer, Heidelberg (2004), http://www.springerlink.com/index/RFT1L2J1UM7W2H86.pdf
16. Glowinski, D., Dael, N., Camurri, A., Volpe, G., Mortillaro, M., Scherer, K.: Toward a minimal representation of affective gestures. IEEE Transactions on Affective Computing 2(2), 106–118 (2011)
17. Winter, D.A.: Biomechanics and motor control of human movement. John Wiley & Sons, Inc., Toronto (1990)
18. Sethares, W.A., Staley, T.W.: Periodicity transforms. IEEE Transactions on Signal Processing 47(11), 2953–2964 (1999)
19. Sneddon, I., McRorie, M., McKeown, G., Hanratty, J.: The belfast induced natural emotion database. IEEE Transactions on Affective Computing 3(1), 32–41 (2012)

Recognizing the Visual Focus of Attention for Human Robot Interaction

Samira Sheikhi[1,2] and Jean-Marc Odobez[1,2]

[1] Idiap Research Institute, Switzerland
[2] École Polytechnique Fédérale de Lausanne, Switzerland

Abstract. We address the recognition of people's visual focus of attention (VFOA), the discrete version of gaze that indicates who is looking at whom or what. As a good indicator of addressee-hood (who speaks to whom, and in particular is a person speaking to the robot) and of people's interest, VFOA is an important cue for supporting dialog modelling in Human-Robot interactions involving multiple persons. In absence of high definition images, we rely on people's head pose to recognize the VFOA. Rather than assuming a fixed mapping between head pose directions and gaze target directions, we investigate models that perform a dynamic (temporal) mapping implicitly accounting for varying body/shoulder orientations of a person over time, as well as unsupervised adaptation. Evaluated on a public dataset and on data recorded with the humanoid robot Nao, the method exhibit better adaptivity and versatility producing equal or better performance than a state-of-the-art approach, while the proposed unsupervised adaptation does not improve results.

Keywords: Human robot interaction, visual focus of attention, gaze, head pose.

1 Introduction

Endowing a humanoid robot with the capacity to interact with multiple persons at the same time requires the design of perceptual algorithms allowing the robot to analyze human behaviors and understand their intent. In particular, it is essential for the robot to be able to recognize communicative behaviors expressed by surrounding people.

In this paper, we addressed the recognition of gaze, and more precisely, the recognition of the VFOA (who is looking at whom or what). VFOA is an important cue for supporting interactions and dialog modeling: it is a good indicator of addresseehood (who speaks to whom, and in particular is a person speaking to the robot), but also a good cue to understand interaction between people or their level of interest. For instance, in a Museum scenario, if people are looking at the painting currently explained by Nao (our project robot), they are probably following the discourse. In order to create effective and natural conversational

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 99–112, 2012.

human-robot interfaces, it is desirable to have robots which can sense a user's gaze and infer appropriate conversational cues [14].

For estimating people's gaze, two main streams of work exist. Active sensing based methodologies based on infrared light are used very often. They are accurate but quite invasive and restrictive [5]. Computer vision techniques on the other hand use perceived information from gaze, head and body posture for recognizing VFOA [12]. This can be done using high definition images of the eyes. Still, it remains relatively constraining and usually restrict the mobility of the subject, considering the need for cameras with narrow field-of-views.

As an alternative, researchers have considered head pose as a clue for gaze [20], [18], [4], [15]. This idea is supported by the fact that turns of the head are a very informative cue in recognizing where the subjects are looking at [12]. Nevertheless, despite being very informative for recognizing VFOA, head pose it is an ambiguous cue: in realistic scenarios, the same head pose can be related to looking at different targets, depending on the situation; conversely, looking at a given target can be done using different head poses, as illustrated in Fig. 2. In this context, the following strategy is often exploited to recognize the VFOA:

- for a given person, track his head and estimate his head pose;
- map the head pose information to VFOA targets (looking at me, i.e. at the robot-, looking at another person, looking down, looking at a painting, elsewhere), and use this information within a recognizer to decode the latent sequence of VFOA targets. Note that the use of other cues like speaking utterance could be exploited as contextual information for recognition [4], [15], but is not addressed in this paper.

In practice, data driven approaches try to directly infer VFOA from head pose without estimating gaze as an intermediate step. Learned parameters, however, are then specific to the geometric configuration between the sensor (robot), the person, and VFOA targets. While this might be suitable in fixed settings [9], it is not adapted for a mobile robot dealing with moving people.

As an alternative we can exploit results from cognitive science studies about human gazing behavior and the dynamics of the head-eye motions involved in saccadic gaze shifts [10,8,11] to automatically determine which head poses should be associated with looking at a given target. This is done using a gaze model relating the head pose, a head-to-gaze ratio, and a head reference direction [3].

This reference direction, which corresponds to the direction perpendicular to the shoulder, was assumed to be fixed in [3] and set according to the setup. This assumption might not hold true in potentially more dynamic settings, e.g. those involving the robot. In these situations, we believe that an explicit or implicit estimation of the reference direction can result in more accurate VFOA recognition. In this context the contributions of the paper are the investigation of two models to dynamically estimate the reference pose, within a VFOA recognition task, and their evaluation on 3 datasets (meeting and robotics domain).

Section 2 goes through the related works. Section 3 reminds the basic Hidden Markov Model (HMM) used to recognize VFOA, the parameter setting issue, and introduce the standard gaze model. In Section 4, we introduce our new models,

providing the intuition behind them and their formal description. Results on meeting benchmark data and on Humavips Nao data are presented in Section 5, while Section 6 concludes the paper.

2 Related Work

In HRI and HCI context, many conversational systems need VFOA information for analyzing and performing necessary interactions. However, with respect to VFOA recognition, most works use either sensor-based or high definition image approaches which are not usually applicable for interaction with robots. The remaining works mostly take a very simplified version of the problem or do not explicitly mention VFOA recognition at all. For instance, in [7], [6] it is not mentioned how the VFOA is extracted and it is only used by the other modules. In [6] the problem is relaxed by only inquiring if the person is looking at the system or not. In [13] detecting a frontal face at a suitable spatial location is enough to adjust the classification to a higher level of engagement. In other works such as it is not clear how they solve the task. In [16] gaze is expressed in terms of head movements but it is not mentioned how to extract it and in [17] it is admitted that gaze is a very fundamental cue in human-human and HRI, but still nothing is mentioned about its extraction.

In another context (meeting), several works explored Dynamic Bayesian Networks (DBN) relying on head pose only [18] or multimodal data [15], [4] for VFOA recognition. All of them rely on Gaussians to model the distribution of head pose for looking at a given target, but only [3] uses a gaze model that does not require annotated data for setting the Gaussian means, or the manual setting of prior values. This allows for an easy exploitation for different observer-VFOA target configuration. Still, the head reference used in [3] is considered to be fixed and and set to the middle of the VFOA targets, it does not evolve dynamically preventing its adaptation to the focus context.

The approach in [19] follows the same approach for setting the means of the Gaussians. In a dynamic scenario they propose to use a discrete set of different head-to-gaze ratios according to the gaze dynamics. From the set of different ratios they take the one with the highest weight for each situation. This is quite different from our proposition since we try to compensate the model limitation by estimating real reference head directions.

3 VFOA Recognition Using HMM

3.1 The HMM Model

A basic solution for inferring the VFOA from head poses is to model the distribution of head poses with a K component Gaussian mixture model where K is the number of existing targets [18]. This method assigns the head poses lying on a specific Gaussian with the corresponding visual target. The model can be easily extended to an HMM as shown in Fig. 1(a), allowing to incorporate temporal information and obtaining more continuous and consistent VFOA results.

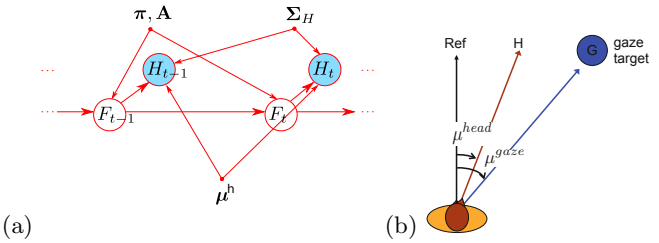


Fig. 1. (a) HMM graphical model for VFOA recognition with raw parameters. (b) Head-Gaze relationship. The person is assumed to be looking at the reference direction at rest (this direction grossly corresponds to the body orientation). Then, looking at a gaze target is accomplished by both the eyes and head. As a first approximation, the head rotation is a linear fraction of the full gazing rotation.

Let H_t and F_t indicate head pose (represented by a pan and tilt angles) and focus values at time t, and A denote the transition matrix in the HMM. Moreover let $\mu^{head} \in \mathbb{R}^{K\times 2}$ and $\Sigma_H \in \mathbb{R}^{K\times 4}$ denote the means and covariances of the K Gaussians. The HMM equations can then be written as follows:

$$P(H_t|F_t = n) = \mathcal{N}(H_t|\mu^{head}(n), \Sigma_H) \tag{1}$$

$$P(F_t = m|F_{t-1} = n) = A_{nm} \tag{2}$$

3.2 The Parameter Setting Issue

A major question is how to set the HMM parameters: the means μ^{head}, covariances Σ_H and transition matrix A. Following previous work, covariances can be set according to the size and proximity and of targets. The transition matrix A can also be set to satisfy our expectation of preserving the continuity in the sequence, an no other preferences. However, setting the means of the Gaussians μ^{head} is not possible in an easy way as it is highly related to the configuration of the observer and the targets and plays the most important role in the model.

The Training Approach. relies on annotated data to estimate the model parameters. However, annotating the VFOA of people in videos is difficult and time consuming, as training data needs to be gathered and annotated for each possible configuration of participant, targets and settings. This is especially problematic if people are free to move.

The Geometric Gaze Modeling Approach and Head Reference Direction. To overcome the above difficulty we can use cognitive findings on gazing behavior [8,11] which state that gazing at a target is accomplished by rotating

both the eyes ('eye-in-head' rotation) and the head (and sometimes even the body in the same direction) as illustrated in Fig. 1(b). The relative contribution of the head and eyes towards a given gaze shift is found to follow simple rules [8], [11]. More precisely, the means of the Gaussians corresponding to each specific target can be set as a fixed linear combination of the target direction and the *head reference* direction. For a gaze target indexed by n, we have:

$$\mu^{head}(n) - R = \alpha\ (\mu(n) - R) \quad if \quad |\mu(n) - R| > \lambda_\alpha \tag{3}$$

or equivalently (if we set λ_α to 0):

$$\mu^{head}(n) = \alpha\ \mu(n) + (1 - \alpha)R \tag{4}$$

where $\mu^{head}(n) - R$ is the rotation made by the head to look at the direction of the target, $R \in \mathbb{R}^2$ denotes the head reference direction and $\mu \in \mathbb{R}^{K\times 2}$ denotes the target directions. The coefficient α is usually set between 0.5 and 0.7 for pan and between 0.3 and 0.5 for the tilt angle. For a given application, a suitable value can be obtained by studying the existing behavior on training data of different individuals.

Equation 4 can be used to set the mean of the Gaussian corresponding to target n in our HMM model. In previous work, the reference vector R was set to a constant value (eg the median of the target directions in [3]). Assuming this as a baseline, the probability of observations given the VFOA states is then given by the following equation:

$$P(H_t|F_t = n, \mu_t) \sim \mathcal{N}(H_t|\ \alpha\mu_t(n) + (1 - \alpha)R,\ \Sigma_H) \tag{5}$$

4 Exploiting Temporal Head Reference Estimates

Setting the means of the Gaussians using the cognitive model requires the knowledge about the value of the reference R as well as the directions of the targets. Equation 4 shows the importance of the reference for recognizing correct targets. Note that using a wrong value for R produces shifted mean values for all of the targets $\mu^{head}(n)$ simultaneously, which can have dramatic effects.

This importance of knowing the head reference is also illustrated in Fig. 2. It shows that, unless the head reference directions (people shoulder's orientation) are more or less constrained by the setting (e.g. like when people are seated in a meeting) or the situation is known (e.g. in the quiz scenario, people are dominantly facing the robot), when can not use a constant reference direction in our model. In more versatile situations and interactions, we have many variations and shifts in the reference as people are free to move. These reasons motivate us to find a suitable way for setting the reference dynamically. Therefore, we proposed two different solutions for setting the reference and their corresponding probabilistic models as explained in the following sections.

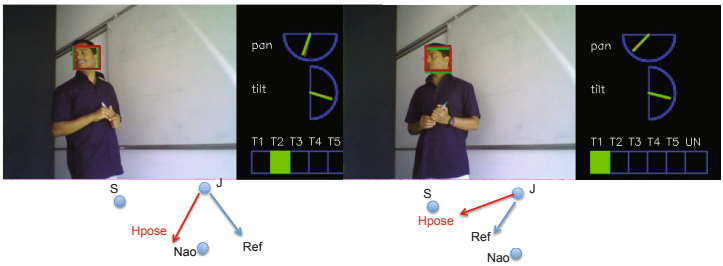


Fig. 2. Different reference directions (shoulder orientations) lead to different poses for looking at the same target. In both images, person J looks at person S. These images illustrate that the geometric model is holding true: the head orientation is approximately half-way between the reference direction and the gaze direction. On the left image, using looking at Nao as reference direction could lead to a wrong interpretation of the head pose on the right as looking at Nao.

4.1 First Model G1

Intuition: For the first model we tend to use a general notion for the reference which is in average acceptable. The principle is that a person tends to orient himself towards the set of gaze targets he/she spends time looking at. Such a body position makes it more comfortable and less energy consuming to rotate his head towards different gaze targets. As a corollary, this means that his average head pose over a time window is a good indicator of his reference direction, and can be used as an estimate of this direction. Although such an estimate might not be very sensitive to local changes and temporal variations and does not account for the previous head pose (that is involved in gaze shifts according to the cognitive studies that led to the geometrical model), it can provide a robust angle estimate that reflects the overal balanced direction of the head or body. Therefore we can set the reference value at each frame R_t^0 to the average of the person's head pose over a previous time window:

$$R_t^0 = \sum_{i=t-w}^{t} H_i/w$$

Setting the reference in this way is also linked to the midline effect [12] which plays an important role in the head direction needed looking at a target.

Model: Since the reference directions set this way are known from the head poses, they serve as the observations in the model as illustrated in Fig. 3. Here again μ_t^{head} denotes the expected means for the head poses. The dynamics between the hidden states are the same as previously and the rest of the relationships are formulated as follows:

$$P(\mu_t^{head}(n)|R_t^0) \sim \mathcal{N}(\mu_t^{head}(n)|\ \alpha\mu(n) + (1-\alpha)R_t^0,\ \Sigma_\mu) \quad (6)$$

$$P(H_t|F_t = n, \mu_t^{head}) \sim \mathcal{N}(H_t|\ \mu_t^{head}(n),\ \Sigma_H) \quad (7)$$

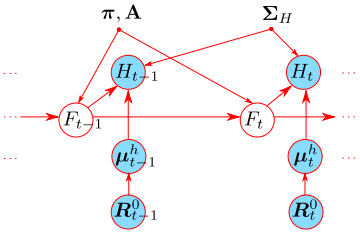


Fig. 3. First model. The head reference direction and the mean head pose of the Gaussians are now variables over time. However, they are observed variables: the head direction is defined as the average of the head poses over a temporal window, and the mean head poses are then deduced from the geometric gaze model.

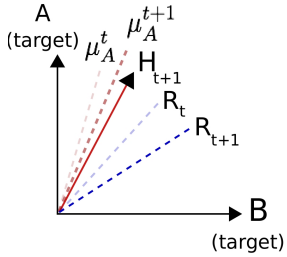


Fig. 4. Unsupervised reference adaptation. Assume that at time $t+1$ the head pose H_{t+1} is associated with target A of current head pose mean μ^t_A in the picture. Trusting the current recognition, adaptation will move the mean μ^{t+1}_A at time $t+1$ closer to the observation and as a result of the gaze geometrical model (assuming there is no change in target positions during this time interval), the reference direction will move accordingly.

The recognition of the VFOA values is straightforward by running the classical inference algorithm on this HMM.

4.2 Second Model G2

Intuition: Setting the reference using long term head pose value statistics might not be sufficient, as more local (short term) gaze activity can come into play. We can thus try another strategy and adapt it in an unsupervised way in order for the model to better fit the observations. This is illustrated and explained in Fig. 4: we would like to change the reference R in order to maximize the probability of observing these new head pose values given their recognized targets.

Model: to accommodate the long term statistics with the short term adaptation, we add a new variable R_t denoting the real reference direction at each frame t and use the estimated head reference direction R^0_t from the average head pose as a prior to this variable. This is illustrated in Fig. 5. By adding R_t as a hidden variable to the model we would then infer the reference values around R^0_t such that the predicted means μ^{head}_t can best fit the observations.

Notice that the same equations 7 and 6 from the previous model are still valid here by substituting R_t for R^0_t. We also expect the reference to be continuous over successive frames and thus the value of R_t should be dependent on its value R_{t-1} at the previous frame. Therefore, we set it as a linear combination of R_{t-1} and R^0_t. The main intention for including the prior value R^0_t in our model is to avoid R from deviating too much from a reasonable range. The following equation formulates this relationship:

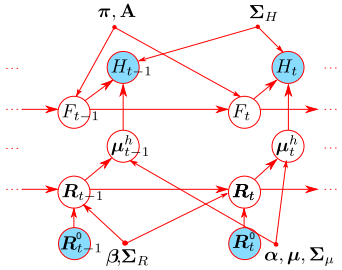


Fig. 5. Second model, unsupervised adaptation for the reference

$$P(R_t|R_{t-1}, R^0) \sim \mathcal{N}(R_t|\ \beta R_{t-1} + (1-\beta)R^0,\ \Sigma_R) \tag{8}$$

Inference: Given the probabilistic model, we wish to determine the sequence of visual focus of attentions F_t (VFOA) of a person from the observed head poses H_t. All the parameters in the model are assumed to be given and remain fixed throughout the inference. We can note that if the head poses μ^{head} are known, our cognitive VFOA model splits into two parts: the VFOA values follow a standard HMM model, whereas the reference variables follow a Kalman filter model. We thus use the following approximate procedure for inference. At time t, we first apply the prediction step for the R_t value and then the targets μ^{head} mean, apply the HMM filtering step for the VFOA state, and then apply the update steps for μ^{head} and R_t given the estimated VFOA for which efficient inference procedures exist.

5 Experimental Results

5.1 Data Sets and Experimental Protocol

For our experiments we use three sets of data. In the first dataset we have the recordings of eight meeting sessions with a total duration of 145 minutes. All of the meetings are recorded under the same condition and with similar configuration as shown in Fig. 6, with four people (Person left Pl and Person right Pr seen on the image, and two organizers O1 and O2 seating in front of them) discussing statements displayed on slides. We perform our study on the two persons on the seats in front of the camera. For this dataset we have ground truth head poses, captured from flock of bird sensors which will be used for analysis. Each of the participants has five possible gaze targets: three other persons, the slide screen and the table.

In the second dataset (D1) we have a video recorded by our robot, Nao. The total duration of this video is 22 minutes. In this case there are two participants seating in front of Nao as shown in Fig. 6. For this dataset, we do not have

Fig. 6. Datasets. (Left) Meeting setting, with VFOA targets for the person on the right (PR). (Middle) Nao dataset 1 (D1) with two participants in front of Nao. (Right) Nao dataset 2 (D2), from Vernissage recordings, with VFOA targets for one of the two participants.

ground truth head poses and analysis are done using the tracked head poses [2] which does joint tracking and head pose estimation [1]. Each of the participants have three visual targets: the other participant, Nao and a booklet which they refer to during the recording.

In the third dataset (D2) comes the Vernissage (the word refers to the preview of an art exhibition) data recording (Fig. 6). There we have one session during which people participate in a quiz given by the robot. The recordings take around 6:30 minutes. Here again we used tracked head poses for analysis. As shown in Fig. 6, there are five main VFOA targets in this recording: Robot (NAO) Person1 (partner), Painting1, Painting2, Painting3. In addition, we use the label Others when people look elsewhere (often down in front of them, with not much head pose change).

Performance Measure: As performance measure we use "Frame based Recognition Rate (FRR)" which corresponds to the percentage of frames during which the VFOA has been correctly recognized.

Algorithms: We have performed our experiments with three different models as summarized in Table 1. The baseline is the basic HMM model using an initially set and fixed reference value for all of the frames. The other models were presented in the previous Section: G1 uses the head pose average over a temporal window as the reference; and G2 adapts the reference value in an unsupervised fashion, using the head pose average value as prior at each frame.

Table 1. Tested algorithms

Model	Reference Prior	Unsupervised Adaptation
Baseline	set as the initial reference	No
Model G1	head pose average over a window	No
Model G2	head pose average over a window	Yes

5.2 Parameter Setting

Meeting Data. We set the variances of the Gaussians according to the size of the targets. For the meeting data we use the same values as in [3] which are $\sigma_\alpha(O_1, O_2, P_R, P_L) = 12$, $\sigma_\alpha(SS) = 25$, and $\sigma_\alpha(TB) = 20$ for the pan, and $\sigma_\beta(O_1, O_2, P_R, P_L) = 12$, $\sigma_\beta(SS, TB) = 15$ for the tilt. Here, σ_α and σ_β show the variances for pan and tilt values. Moreover, O_1, O_2 indicate the observers, P_R, P_L the persons on the right and left, and SS, TB indicate the slide screen and the table respectively.

For the gaze directions, they were assumed to be fixed for each recording (thus neglecting people's motion), and currently defined from the geometrical setting. The initial value for the reference direction is particularly important for the baseline for which it remains the same over time, but not important for the other models as the reference value is quickly set as the average over head pose values. For the baseline, we experimented with setting the reference as the middle of the gaze target directions, which was shown to work the best in previous works [3].

Table 2. Parameter set using cross-validation

Parameters	Baseline	1st Model G1	2nd Model G2
α_{pan}	$\surd$	$\surd$	$\surd$
α_{tilt}	$\surd$	$\surd$	$\surd$
self-loop	$\surd$	$\surd$	$\surd$
window-size	$\times$	$\surd$	$\surd$
σ_R	$\times$	$\times$	$\surd$
ratio σ_μ/σ_H	$\times$	$\times$	$\surd$
β	$\times$	$\times$	$\surd$

The remaining parameters of the models which are summarized in Table 2 were adjusted by cross-validation separately for each of the models. For the meeting data there are two different set-ups for people seating on the first and second seats and therefore we did the cross validation once by taking the training set from the same seat and once by taking it from the other seat. The second case is useful to evaluate whether our model is sensitive to a specific setting or it is more general. For training with data from the same seat, leave-one-out cross validation was used, taking seven meetings for training and testing on the eighth meeting. For training with data from the other seat, all eight meetings from the other seat were used for training and obtained parameters were used to test on the other seat.

Table 3 summarizes the parameters which were obtained in cross validation for the baseline and Table 4 summarizes chosen parameters for the first model G1. There is a strong agreement between the parameters which were obtained for left and right people. Also there is a strong overall consistency between the

parameters trained using the first seat data and those of the second seat. For the baseline model, α_{pan} obtained from two different seats is different. This could be through to the fact that the reference which is used there (the middle of the targets) is a poor reference and force and introduces different results for these two different settings. This effect does not exist for the fist model G1 and the chosen parameters are completely consistent.

Table 3. Chosen parameters for the meeting data and baseline model

Person	Training	Parameters: α_{pan}, α_{tilt}, self-loop
Person on left	same seat	0.5 - 0.5 - 0.75
Person on left	other seat	0.8 - 0.5 - 0.75
Person on right	same seat	0.8 - 0.5 - 0.75
Person on right	other seat	0.5 - 0.5 - 0.75

Table 4. Chosen parameters for the meeting data and first model G1

Person	Training	Parameters: α_{pan}, α_{tilt}, self-loop, wind-size
Person on left	same seat	0.7 - 0.5 - 0.75 - 500
Person on left	other seat	0.7 - 0.4/0.45 - 0.75 - 500
Person on right	same seat	0.7 - 0.4/0.45 - 0.75 - 500
Person on right	other seat	0.7 - 0.5 - 0.75 - 500

Nao First Dataset (D1). For the gaze directions, same as the meeting data, they are assumed to be fixed for each recording and defined from the geometrical setting. The initial value for the reference direction is considered to be at Nao's direction which is a reasonable choice in human robot interaction scenario. We set the standard deviations of the targets to 8° for the pan and 4° for the tilt angle. Notice that these values are smaller compared to the meeting data. This choice is made both our Nao datasets where we use tracker results for head poses since those head poses are usually smaller than the ground truth pose values.

For the rest of the parameters, as there are a few number of people participating in this dataset with very different gazing behaviors cross-validation will not produce reliable parameters. To choose the parameters we consider the meeting data as the training set and use parameters obtained from that data for running our algorithms on Nao's data. Note however, that the resulted α_{pan} value from meeting data is 0.7. In Nao data this ratio is big considering the tracked head poses which are a little underestimated. Therefore we do our experiments with a smaller value of 0.65.

Nao Second Dataset (D2). Standard deviations of the pan angles were set to 8°, 8°, 8°, 9°, 10° respectively for Robot, Partner, Painting1, Painting2 and

Painting3, according to their size and proximity. The tilt angle standard deviations were set to 4° for all targets. The remaining parameters are all set in the same way they are set for D1.

5.3 Results

Meeting Data. Table 5 shows the results of the baseline, G1 and G2 models. As can be seen, the first model outperforms the baseline. This is particularly true in more mismatched conditions, when parameters are learned from another seat rather than the same seat, thus exhibiting a better adaptation capacity. In particular, we can notice the performance degradation for person right (PR) when using the optimal parameters for person on left (PL). The main (mismatched) parameters leading to the degradation is the parameter α of the gaze model (see Eq. 4) that directly impact the prediction of the head poses: for PL, the optimal parameters is around 0.8, whereas for PR, it is around 0.5. Using the head pose average for the reference is indeed a more stable choice for this important parameter, with an optimal value for both seats around 0.7.

On the other hand, we can see that the 2nd model G2 performs very closely to the G1 model, a behavior that will be seen in other recordings as well. This means that in practice the unsupervised adapted head reference remains very close to the prior, and thus the models behave very similarly.

Table 5. Performance evaluation on Meeting data

Person	Training	Baseline	Model G1	Model G2
Person on left	same seat	64.7	**65.7**	65.5
Person on left	other seat	64.5	66.7	**66.8**
Person on right	same seat	57.0	**58.7**	58.6
Person on right	other seat	43.9	**59.0**	59.0

Nao First Data D1. The results of the baseline, G1 and G2 are summarized in Table 6. Despite the quite different setting (situation, number of gaze targets, use of estimated head pose vs ground truth head pose), the conclusions are similar to the meeting data. More precisely, model G1 outperforms the baseline, particularly for people on the left with a large difference, and model G2 performs almost the same as model G1.

Table 6. Performance evaluation on Nao data D1

Person	Baseline	Model G1	Model G2
Person on left	69.4	**78.7**	**78.8**
Person on right	**66.5**	66.2	66.2

As the table shows, we have a high gain using our models for the person on the left while the performance of the models are very close for the person on

the right. The main reason for this behavior is that considering the participants body configuration Nao's direction is quite a suitable choice for the reference for the person on the right but it cannot perform as a good reference estimation for the person on the left.

Nao second Data D2. Table 7 shows VFOA recognition rates obtained our first model compared to the baseline model. As it is shown in the table for person on the right we get better results using our model G1 whereas for the person on the left this is not true. We need to verify these results using more sequences from this dataset to find the overall performance behavior.

Table 7. Comparison of FRR of Baseline and the 1st model G1 on quiz recording 9

Person	Baseline	Model G1
Person on Left	**54.6**	51.2
Person on Right	58.8	**59.6**

6 Conclusion

In this Section, we have presented our research towards designing better gaze models for improved VFOA recognition. We have shown that the implicit estimation of the head reference has a positive impact on performance. It is important to underline that for the different recordings the choice of head reference for the baseline is a very good approximation of the actual value. In practice, such a value might be difficult to set. As we have seen, in the robot interaction application, the same strategy (looking at Nao) does not produce good results in all conditions. Similarly, using the middle of the gaze target directions might work, but assumes that the robot is aware of all target directions a person can be looking at. This might not hold true in all cases.

Our future work contains an assessment of the model on larger amounts of recordings with the robot from the Vernissage dataset; assessment of the model using the true head poses, to mitigate the effect of head pose estimation on performance evaluation. Moreover, we would compare the results of our VFOA models using tracked head poses versus the ground truth head poses to study how the head pose estimation error affects the VFOA recognition.

Acknowledgments. The authors gratefully acknowledge the financial support from the HUMAVIPS project, funded by the European Commission Seventh Framework Programme, Theme Cognitive Systems and Robotics, Grant agreement no. 247525.

References

1. Ba, S.O., Odobez, J.-M.: Evaluation of multiple cue head pose estimation algorithms in natural environements. In: IEEE Int. Conf. on Multimedia and Expo (2005)

2. Ba, S.O., Odobez, J.-M.: Probabilistic Head Pose Tracking Evaluation in Single and Multiple Camera Setups. In: Stiefelhagen, R., Bowers, R., Fiscus, J.G. (eds.) RT 2007 and CLEAR 2007. LNCS, vol. 4625, pp. 276–286. Springer, Heidelberg (2008)
3. Ba, S.O., Odobez, J.-M.: Recognizing visual focus of attention from head pose in natural meetings. Trans. Sys. Man Cyber. Part B 39, 16–33 (2009)
4. Ba, S.O., Odobez, J.-M.: Multiperson visual focus of attention from head pose and meeting contextual cues. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 101–116 (2011)
5. Babcock, J.S., Pelz, J.B.: Building a lightweight eyetracking headgear. In: Proceedings of the 2004 Symposium on Eye Tracking Research & Applications, ETRA 2004, pp. 109–114. ACM, New York (2004)
6. Bohus, D., Horvitz, E.: Models for multiparty engagement in open-world dialog. In: Proc. of the SIGDIAL Conference, Stroudsburg, USA, pp. 225–234 (2009)
7. Bohus, D., Horvitz, E.: Open-world dialog: Challenges, directions, and prototype. In: Proceedings of IJCAI 2009 Workshop on Knowledge and Reasoning in Practical Dialogue Systems (2009)
8. Freedman, E.G., Sparks, D.L.: Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. Journal of Neurophysiology 77(5), 2328–2348 (1997)
9. Gaschler, A., Huth, K., Giuliani, M., Kessler, I., de Ruiter, J., Knoll, A.: Modelling state of interaction from head poses for social human-robot interaction
10. Hanes, D.A., McCollum, G.: Variables contributing to the coordination of rapid eye/head gaze shifts. Biol. Cybern. 94, 300–324 (2006)
11. Hayhoe, M., Ballard, D.: Eye movements in natural behavior. Trends in Cognitive Sciences 9(4), 188–194 (2005)
12. Langton, S.R., Watt, R.J., Bruce, I.: Do the eyes have it? cues to the direction of social attention. Trends Cogn. Sci. 4(2), 50–59 (2000)
13. Michalowski, M.P., Sabanovic, S., Simmons, R.: A spatial model of engagement for a social robot. In: 9th IEEE Int. Workshop on Advanced Motion Control (2006)
14. Morency, L.-P., Darrell, T.: Conditional Sequence Model for Context-Based Recognition of Gaze Aversion. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 11–23. Springer, Heidelberg (2008)
15. Otsuka, K., Takemae, Y., Yamato, J.: A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In: Proceedings of the 7th International Conference on Multimodal Interfaces, ICMI 2005, pp. 191–198. ACM, New York (2005)
16. Sidner, C.L., Lee, C.: Engagement rules for human-robot collaborative interactions. In: IEEE Int. Conf. on Systems, Man and Cybernetics, vol. 4 (2003)
17. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artificial Intelligence 166(1), 140–164 (2005)
18. Stiefelhagen, R.: Tracking focus of attention in meetings. In: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, ICMI 2002, p. 273. IEEE Computer Society, Washington, DC (2002)
19. Voit, M., Stiefelhagen, R.: Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In: Proc. of the 10th Int. Conf. on Multimodal interfaces (ICMI), Chania, Crete, Greece (2008)
20. Yücel, Z., Salah, A.A.: Resolution of focus of attention using gaze direction estimation and saliency computation. In: Proceedings of the International Conference on Affective Computing and Intelligent Interfaces (2009)

Contextual Analysis of Human Non-verbal Guide Behaviors to Inform the Development of FROG, the Fun Robotic Outdoor Guide

Daphne E. Karreman, Elisabeth M.A.G. van Dijk, and Vanessa Evers

Human Media Interaction, University of Twente
Enschede, The Netherlands
{d.e.karreman,e.m.a.g.vandijk,v.evers}@utwente.nl

Abstract. This paper reports the first step in a series of studies to design the interaction behaviors of an outdoor robotic guide. We describe and report the use case development carried out to identify effective human tour guide behaviors. In this paper we focus on non-verbal communication cues in gaze, gestures and movements. The work reported involves the observation of human tour guide behaviors and visitor responses as well as interviews with guides. An affinity diagram is used to identify effective communication cues of human guides and the relations between them. The opportunities for a robotic guide are discussed. We argue that human guide behaviors and strategies cannot be one-on-one applied to robot tour guides. Instead, we aim to develop abstractions of the human behaviors, appropriate for robot tour guides and effective in realizing visitor engagement. The results of this study will be used to create a first Fun Robotic Outdoor Guide prototype with the abstracted interactive robot guide behaviors implemented to assess the effects on visitor experience in 'the wild.'

Keywords: Human Tour Guide Behavior, Non-Verbal Robot Behavior, Contextual Analysis.

1 Introduction

The EU 7th Framework project FROG (Fun Robotic Outdoor Guide www.FROGrobot.eu) aims to develop a guide robot with a winning personality and behaviors that will engage tourists in a fun exploration of outdoor attractions. The work involves innovation in the areas of vision-based detection, robotics design and navigation, human-robot interaction, affective computing, intelligent agent architecture and dependable autonomous outdoor robot operation. In this paper the focus is on the Human-Robot Interaction (HRI).

In many museums and tourist sites human tour guides are guiding visitors and convey information about the sites. However, not all visitors want to join or cannot afford a two hour guided tour. We believe that an autonomously navigating mobile robot can provide information to visitors in new and innovative ways.

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 113–124, 2012.

The first step in developing the FROG-robot was to determine the requirements for the robot in the different tourist sites and identify opportunities to improve the visitor experience with a robotic guide. From this analysis it became clear that the FROG-robot needs to be available for small groups of visitors, will present some interesting and historical information and curiosities about the site and will guide visitors for a limited amount of time through a part of the site. The information conveyed will be based on the visitors interests [1].

The goal of the HRI part of the FROG project is to determine personality and behavior for the robot. We argue that using anthropomorphic communication cues will help visitors understand the robot, because humans naturally respond to non-verbal communication cues. The next step in the development of the FROG-robot, described in this paper, is exploring the non-verbal robotic guide behavior, such as gaze behavior, gestures and movements. This will be done by first observing and analyzing the human tour guide behavior, strategies and personality. Possibilities and limitations of transferring this behavior to robots will be examined based on literature.

In Section 2 the related work on robotic guide and communication behavior and human tour guide behavior is presented. The methodology of observing and analyzing behaviors of human tour guides and the visitor responses are given in Section 3 and subsequently the results of the analysis are reported in Section 4. Suggestions on how to apply the human behavior cues to a robot guide are indicated in Section 5. Finally, conclusions and future work are presented.

2 Related Work

Research on robotic museum guides has focused on various aspects. The robot Minerva, does quite a good job on interaction with visitors, because this robot is able to express itself in different moods [2]. The Robovie robot in the science museum does very well in addressing visitors and keeping their attention [3]. However, the interactions between the robots and the humans are still limited, because humans are still testing the boundaries of the systems [4], or are distracted by seeing a robot and lose interest in the exhibit.

Research has also looked into the influence of different modalities, e.g. the gaze of robots has been proven to be very important. As Mutlu et al. [5] found, robot gaze behavior influences the abilities of visitors to recall a story. When the robot looked at the visitors, the listeners could remember the story better than when the robot was looking around randomly [5]. Also gaze behavior in combination with pointing seem to be very important, because human tour guides' use of head movements at communication relevant places helps humans understanding the story told. Kuno et al. [6] tested these behaviors in a robot and the head movements were efficient for understanding the robot, too.

The robots mentioned above are (more or less) based on behaviors human tour guides show. Humans use effective guide behavior, so looking at the behavior of human tour guides may help setting the behavior of robotic tour guides. As Duffy states: using anthropomorphism and anthropomorphic communication cues can be powerful

and intuitive to make humans understand and naturally interact with robots [7]. We have to keep in mind that humans respond more rapidly to humans than to robots as Kanda et al. [8] found, because this will influence the human robot interaction.

However, the robots in these examples are only doing parts of the tour guiding, e.g. telling the story, but not interacting with human and conveying information very well at one exhibit, but cannot perform at another. Finally joining a robot tour guide should be as satisfactory as joining a human tour guide. So to determine effective behaviors for robot guides, it might be helpful to study effective human tour guide behaviors, strategies and procedures.

Some related work has been conducted in this direction with human tour guides. This work found that originally tours were like lectures and were more or less a monologue of the tour guide [9], but nowadays the tours have become more adjusted to the interests of the visitors which has the advantage that visitors are more involved in the tour and like the tour better [10]. To adapt the tour to the interest of the visitors, the guide needs to be able to tell flexibly about everything they encounter, so visitors do not notice the change in the tour when the guide makes changes in content, e.g. in case of some places of the trip not being available [11].

To give an engaging tour, guides use several strategies to get and keep the attention of the visitors. One of these strategies is to interact with them [9]. The interaction with visitors (e.g. verbal interaction) is important to keep the visitors' attention. And except from giving cues to the visitors, the guides also obtain a lot of non-verbal feedback about involvement of the visitors by looking at them. Visitors that are gazing at the guide or the object of interest and who are nodding or smiling are interested, the ones looking away or talking to each other may not interested anymore [10].

Robot tour guides as well as human tour guides should have behaviors and strategies that modify the tour in a positive way. For determining the robot behaviors, looking at human tour guide behavior might be helpful, as proven by the examples above.

3 Methodology

To get insight in the effective human guide behavior, especially gaze behavior, gestures and movements, a qualitative research was performed by two researchers following four guides guiding different groups in two touristic sites. Notes taken during the tour, answers of guides in the interviews, notes from analysis of video data and notes of literature search were combined in analysis using an affinity diagram. The result of the research is a connection diagram and in-depth analysis of human tour guide behavior.

3.1 Research Context

For the analyses two different outdoor tourist sites with guided tours were selected. These sites were the Lisbon City Zoo in Lisbon, Portugal and the Royal Alcazar in Seville, Spain. Both sites offer interesting and challenging opportunities for having robot guides guiding visitors. In these sites the amount of information that is available

without a guide is not satisfying the visitors [1] and these sites offer outside environment to be covered by the FROG-robot.

The Lisbon City Zoo is a park showing several species of wild animals to humans and educating the visitors about nature and animals. Besides that, the Zoo also provides access for scientific research and participates in conservation programs for species. The guides giving tours in the Lisbon City zoo are educated and employed by the Zoo. Visitors of the Zoo are mainly families with one or two (young) children, couples with or without children, school classes and groups of friends. The day is experienced as a social day out.

The Royal Alcazar is a royal home, the first building was built in the ninth century and during ages Christians and Muslims built, destroyed and rebuilt the buildings in the site. The guides in the Royal Alcazar are educated and employed by different agencies or entrepreneurs. All guides must have certification, but the board of the Royal Alcazar does not control the guides a lot. The guides that contributed to this research were certificated. Visitors of the Royal Alcazar are mostly couples (with older children), groups of tourists and school classes. The purpose of the visit is to learn about the history of the site.

3.2 Sample

For the research a total of four guides were observed, video-taped and interviewed. In the Lisbon City Zoo two guides participated. The first guide (male, ten years of experience) guided a group of seven adult visitors. The second guide (female, some years of experience) guided a school class of 19 children aged 9-10 years old. In the Royal Alcazar two guides participated, both were female. The first (ten years of experience) guided a group of eight adult persons, and the last guide (several years of experience) guided a group of twelve adults (and two small kids).

3.3 Procedure

At the start of the tour all visitors and the guide were informed about the research and the filming. In the end of the tour the group was lead to a room and all adults completed a consent form.

Two researches joined the four tours. One of the researchers was video-taping the whole tour. In the tape the guide and his/her expressions and some visitors or the whole group of visitors are visible. The story is not always clear hearable on the tape, but non-verbal behavior of the guide and the visitors' responds is very well visible.

The second researcher followed the guide close and made notes on the story the guide told, outstanding guide behavior and on the events that happened during the tour.

After the tour the researchers interviewed the guides and took individually notes of the answers. When the guide had left, the researchers completed the notes of the observation of the tour and the interview.

3.4 Measures

Notes were taken during and completed after the tour on things that stood out about guide behavior. One of the researchers was following the guide close and took notes during the tour, the second researcher was filming and took the notes at the end of the tour.

After the tour, the two researchers had a short semi-structured interview (approx. 15 minutes) with the guides about the tour they just gave, their experiences guiding different kind of groups, use of strategies and how they would like to improve the visitor experience in the sites. (What is the purpose of your tour? / What is the main exhibit in the site you are guiding? / Do you notice differences between groups? How do you deal with that? / Is guiding children different form guiding adults? If yes, in what way? / How do you get and keep attention of the visitors? / What do you want to change to improve the visitor experience?).

All tours were videotaped to later look back at the actions of the guides and take notes of more specific actions and behaviors. Next to the global analysis, from all guides two film fragments (approx. 2-4 minutes) were taken to analyze in detail. Aspects that were analyzed were: the orientation of the guide, the story guides were telling, the movement and gestures the guides made, to what the guides were looking and how they ended at exhibits.

3.5 Data Analysis

The data gathered from the different researches was combined in the analysis using an affinity diagram. This method is based on Grounded Theory method; themes and results of the research emerged from the data. Using an affinity diagram is very useful when large amounts of qualitative data has to be analyzed, from which the results are complex and not easy to grab [12]. The affinity diagram helps to order the information and to find logic and natural relationships between the parts of the data.

Globally the method follows a few steps. First statements have to be written on cards (post-its or index-cards). These statements are remarkable in a certain way and are taken from all parts of the research. When writing down these statements, no attention should be given to duplications of text or solitary cards. For interpretation of the data color coding of the statements could be useful.

Second, all statements will be shuffled and pasted on a wall. Cards with similar statements will pasted close to each other, but also other relations between statements can be made clear. In this phase clusters of similar subjects and relations between subjects appear. Important is that a statement can be in more clusters (by duplicating the card) and relations can be of all kind (e.g. cause, opposite, similar).

In the third step the clusters will be named and the relations between the clusters will get meaning. Some extra cards with the main findings per cluster will be placed to make the affinity diagram more easily readable. When the number of clusters is high, some can be combined in one larger cluster having some sub clusters. But also the other way around, when the clusters are very large, some sub-clusters can be made.

For the research on human tour guide behavior, from all research methods (observation, interview, video analysis and literature) statements and observations were taken and written on small cards. The statements were color coded by resource (e.g. all statements taken from the first observation were written in blue and statements taken from the interviews were written in purple).

When all cards were completed, the researcher started to cluster the statements on a large wall. During this clustering the placement of the cards was not fixed and cards were removed and replaced if necessary. Finally, all cards were pasted on the wall and names for the clusters were invented, some clusters with sub-clusters were invented. The main statements per cluster were added on yellow notes.

The final affinity diagram was quite large and to make it easier to read, a connection diagram based on the affinity diagram was made in which the clusters and sub-clusters and the relations between them are visible. These connections were counted. And five collections of clusters and sub clusters were found. In the result section the results for the analysis are given, and more explanations about results are presented in the connection diagram.

Please note that the statement cards and the affinity diagram were made by one of the researchers. However, the researchers were discussing the notes on the observations and the video analysis before writing the statement cards. After making the affinity diagram and the connection diagram the researchers discussed the results and added relations, connections and changed names for the clusters to get the best overview of the results.

4 Results

In the connection diagram (Fig. 1) the names of all found clusters are given and the clusters are connected with lines when there was some relation between them (e.g. the connection between distraction and taking pictures is that when visitors get distracted during the tour, they often start to take pictures or, vice versa, they get distracted because they take pictures). The clusters and sub-clusters with the most connections are highlighted in yellow. These are main aspects during a tour and the guides tried to do everything to influence these factors positively.

4.1 Clusters

Ten clusters with sub-clusters appeared, which are briefly described (detailed description and explanation of non-verbal communication cues follow in the next paragraph):

Attention. Without getting and keeping the visitors' attention a tour guide cannot give a tour. Guides used lots of strategies to get and keep the visitor's attention, such as interacting with the visitors and breaking eye-contact and starting to move in the next direction at the end of an exhibit.

Interaction is usual in tours these days compared with the past. All guides knew that a tour of two hours was too long for visitors to just listen to a guide. Therefore, the guides tried to interact with the visitors by addressing them in different ways, such as

asking and answering questions, showing visuals, letting visitors experience (touch, smell) the site, pointing at objects, searching for differences together and by asking for personal interests and referring to that later and having a chat with the visitors during the walk.

Information about the site is conveyed by the guide. The guide usually can tell flexibly about everything they encounter in the site, and answer all the questions of the visitors. Particularly, visitors like to hear curiosities.

Adaptation of the tour is always done to fit the tour to the different type of visitors. But also during the tour to keep visitors interested guides adjust the tour to the group. The guides try to shape the tour around the visitors interests, because when adding something new to a subject visitors were already interested in, it was easy for the guide to keep the attention. Adaptation can be in content, speed or route.

Taking pictures was something lots of visitors did when they were in an attractive place. Taking pictures either meant they were really interested, taking pictures of the subject of interest. Or it meant the visitors were distracted and they started walking around and taking pictures of everything. The visitors that were taking pictures during the tour were often lost and missed much of the story the guide was telling.

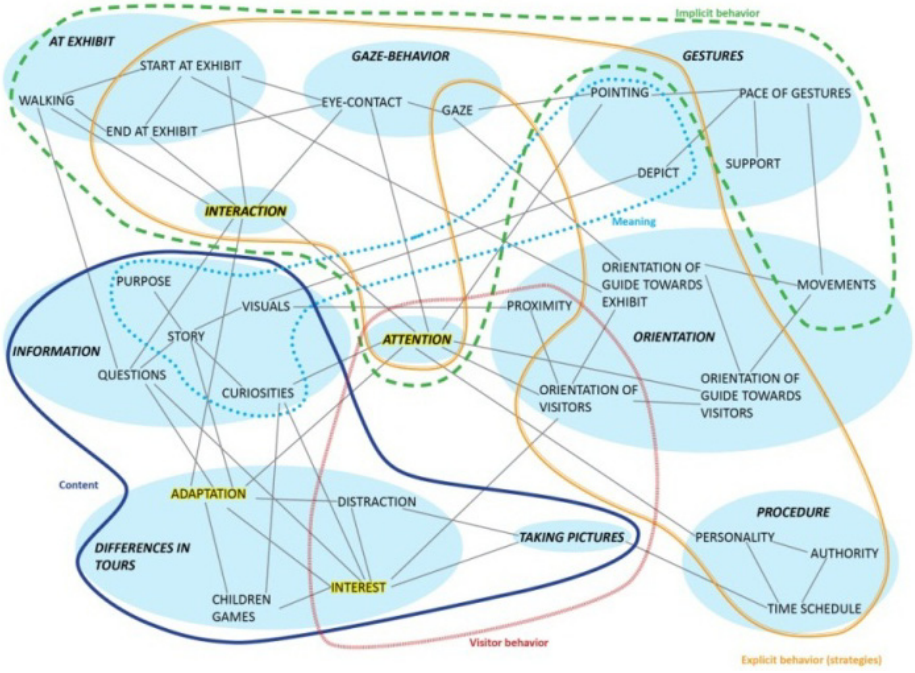


Fig. 1. Connection diagram of Human Tour Guide Behavior

The procedure of the tour is determined by the tour guide, following a tight schedule. The guides were not using authority for adults, because they could choose themselves whether they wanted to listen. The content was based on the visitor interests.

Orientation is the way guides orientated towards visitors and to the exhibit; always in a way visitors were able to see both the exhibit and the guide.
Gestures made by the guide might have helped the guide telling the story or helped the visitor understand the story. All guides made a lot of arm gestures, varying from supporting or depicting the story and pointing.

Gaze-behavior is quite intuitive, but guides used gaze to keep attention by alternating their gaze between visitors, and they were able to see from the visitors' gaze if they were interested. The guides adjusted the story to the ones they were looking at.

At the exhibits guides did start and finish with less important sentences, so all visitors were attended. To keep the tour going, the guides did not wait until the group was complete. They started to talk to the nearby visitors.

4.2 Effective non-Verbal Behavior

Given the page limit and the scope of the workshop this paper only describes non-verbal behavior, because this will give meaningful insights in the possibilities for the FROG-robot. The behaviors were classified as explicit behavior, or strategies the guides used, and implicit behavior. Explicit behavior is behavior the guides employed consciously to attract and keep the visitors attention. The guides were very different people, however, the strategies they adopted showed several commonalities. However, humans are sometimes not aware of non-verbal communication cues, still these cues could be very helpful for the communication. These cues used implicitly by the guides are also described below. Using strategies perfectly fitted in the guides' normal behaviors. Acting natural and naturally modify the tour in a positive way is what the guides wanted to achieve.

Starting and Closing the Story

At the new exhibit the guide never started with the main story. To get the attention of the visitors at the start of an exhibit the guide started the story with some less relevant words or sentences. These can be "ok," or "so," or "this is a very beautiful view." When the group of visitors was not complete yet, the guides did start anyway to keep pace in the tour. When the guides wanted to tell something important and not all visitors were close, the guides raised their voices. The guides concluded with a short summary, telling where to go next or with words like "ok" "so" "let's go" to indicate they were finished at an exhibit.

Gaze-Behavior

Gaze-behavior of the guide is important for the visitors' understanding of the story, but also the other way around. The guide obtains feedback from the gaze behavior of the visitors.

The guides reported in the interviews they alternated their eye-contact evenly between the visitors, as a strategy to keep their attention. The guides indeed were alternating their gaze, but in our analysis of the video material, the guides seemed to choose one visitor at each exhibit to talk to. Most of the time this visitor got the attention and the guide sometimes shifted attention to other visitors, but always got back to the chosen one. However, for different exhibits, the guides seemed to choose different visitors to address. The visitor the guide was looking at was often nodding, turning its head towards the exhibit and back and looked to the guide. This provided the guide with information about engagement of the group.

When pointing at an object of interest, the guides also looked at the exhibit for a while. For the guide, this was to check where to direct the group attention, but sometimes visitors reacted to it with looking into the same direction, following in mutual gaze. When the guides made the visitors look at an exhibit or visual, the guides decided when to go on with the story. Most of the time, the guides waited for most visitors to indicate they had seen it (by nodding or gazing at the guide again). To keep the story going, the guides did not wait for all visitors to look back. On the other hand, the guides sometimes provided extra time to look at something for a particular visitor, by adding non-important sentences to fill the time.

When the guides wanted to go to the next exhibit, they broke eye-contact with the group of visitors or focused their attention to another object or to the visitors in a social way and made a move towards the next exhibit. For visitors that made clear the story was over. And because the guide had moved slightly into the new direction during the last sentence, the visitors knew where to go for the next exhibit.

Gestures and Movements

All guides used a lot of gestures, which can be categorized in pointing, depicting the story and supporting the story. While telling the story all guides used their arms to depict and point to the subject. The depicting of the story and pointing at objects is explicit behavior, the guide knows how, why and at which moment to perform the action in support of the story.

The visual support could be in the site itself, and the guide pointing and touching it, the guide could show visuals, and the guide was able to depict the story if the subject they were telling about was not visible at the moment.

Depicting the story and pointing to exhibits helps the visitor to understand the story the guide is telling. The guide will only depict parts of the story if the subject is not visible at the moment. Otherwise the guide will point to the exhibit to make clear what he/she is talking about.

When the guides walked away from an exhibit, they slightly moved in the next direction during the last sentence of the story. This made clear for the visitors the story was finished and which direction to go next. Sometimes the guides made a follow-me sign, and always the visitors followed like a chain reaction, the nearby visitors starting to walk first and the furthest visitors following last.

When looking at the pace of the gestures and the movements the guides made, except for using arm gestures there are no commonalities. These gestures and

movements were made unconsciously and might help the guide telling the story, at the moment is still not determined what influence they have on the transfer of knowledge. These gestures and movements were personal for each guide and fitted with the overall personality of the guides.

5 Discussion

The robot for the FROG-project is not envisioned to look very human-like, and simply copying the human communication cues and applying them on the robot may not be effective. However, using anthropomorphic aspects of appearance and behavior will be effective. How to design these anthropomorphic aspects, and what communication cues the robot will use depends on the human understanding of the appearance and adopted cues.

The FROG-robot will be approximately 1.20-1.50 meters high, smaller than an average adult to not scare the visitors. The robot will probably not have arms, because for a large not human-like robot unexpected moving parts can scare visitors or even worse, harm them. The next series of researches will give insight in the possibilities in pointing for a robot without arms. Now our intention is to abstract the different communication cues as much as possible, but still keep them intuitively understandable for humans interacting with the robot.

Explicit behavior and strategies the human tour guides showed, such as telling curiosities, addressing the visitors, breaking eye contact and moving in next direction during the last sentence, should be adopted for the robot. These strong cues in human-human interaction have to be translated and abstracted to fit the robot appearance and personality. So the strategies of a human tour guide might not have to be copied one-on-one to the robot, but the results in visitor behavior should be comparable.

From the implicit behavior it is important to look at the gaze behavior; the guides were alternating their gaze between the visitors, but tend to choose one visitor per exhibit to talk most to. When alternating their gaze, they always turned back to that one visitor, who gave the most (implicit) feedback. The visitors chosen were different per exhibit. Further tests should prove if this is important for communication.

The robot behavior when starting or closing a story will be similar to human behavior. The FROG-robot should get the attention at the new exhibit again, and therefore will not start with the most important information, but with a less important sentence. In the end the robot will break “eye-contact” to indicate the story at the exhibit is finished. Also starting to move to the next destination during the last sentence at an exhibit will be of importance for the robot to communicate to the group of visitors where to go next.

Also gaze is very important because humans use a lot of making and breaking eye-contact in human-human interaction. In human-robot interaction human communication cues applied to robotic guides help visitors to understand and recall the story [5, 6]. In the design of the robot gaze behaviors, examining the effective human gaze behaviors and translating them to effective robot guide behavior will probably help the robot conveying information.

Similar like human tour guides the FROG-robot will choose one of the visitors to get the main feedback from. Probably this will be the closest visitor, because the face would be easiest to track. This visitor will give information on the level of interest, by the robot examining the gaze direction, nodding and laughing. Further research will give insight in the necessity of alternating the robot gaze. For different exhibits, the robot will focus on different visitors, as not always the same visitor will be closest.

Looking into the exhibit when talking about a point of interest will be of big importance for the robot than for human. Because the robot will probably not be able to point with arms, other options will be considered. Looking at the exhibit is one of them (next to laser/projecting on the wall). Other than humans, the robot can look to the exhibit for longer period of time, because the "eyes" are not necessary the cameras that are examining the faces of the visitors.

Human tour guides use a lot of gestures, but the robot will not be able to do so. The function of these supporting gestures is not determined from the video-analysis. Depicting a story is important, but the robot will also not be able to depict the story with arms. The robot will have different applications to depict the story, by showing a film on a (touch-) screen, or projecting information on a wall. Also the visuals a human tour guide use, can be showed on screen or projected on the wall by the robot.

The FROG-robot will use some strong anthropomorphic appearance and communication cues to make the interaction with humans intuitive. To achieve this, effective human communication cues should be translated to fit the robot appearance. The non-verbal communication cues used for the robot will be the gaze behavior when talking about a point of interest, breaking eye-contact in the end, starting to move during the last sentence at and exhibit and showing visuals on screen or projection on the wall. Using these human communication cues in abstracted form should have the same result on visitor behavior and attention as the cues a human tour guide is using.

6 Conclusions and Future Work

Robots are more and more successfully used in human social environments. FROG aims to develop a Fun Robotic Tour Guide to guide visitors in indoor and outdoor museums and touristic sites.

This paper presents an analysis of non-verbal human tour guide behavior. The affinity diagram used was very appropriate to use for the analysis of the large amount of qualitative data from varying sources. The method led to a connection diagram, which presented the main aspects of a guided tour and the guide behaviors. To have insight in the behaviors collections were made. This paper focused on non-verbal behavior.

Strategies (explicit behavior) a human tour guide uses can be copied to a robot. Comparison with literature and own tests will give insight in how to use these strategies to achieve the same effect on the visitor behavior.

Implicit behavior of human tour guides includes behavior that helps the guide to tell the story, but it contains also behavior that helps the interaction with visitors. This behavior, like breaking eye contact and already moving in the new direction of the next exhibit will be very useful for the robotic guide.

To conclude, the observed human tour guide behavior such as gaze, showing visuals and movements will be translated to robot behavior and applied on a robot. In controlled lab-experiments, possible robot gaze, movements and orientation will be tested and evaluated on effectiveness, efficiency, understanding and experience of humans.

Acknowledgements. The research leading to these results received funding from the European Community's 7th Framework Programme under Grant agreement 288235 (http://www.frogrobot.eu/).

References

1. Karreman, D.E., van Dijk, E.M.A.G., Evers, V.: Using the visitor experiences for mapping the possibilities of implementing a robotic guide in outdoor sites. In: The 21th IEEE International Symposiums on Robot and Human Interactive Communication, ROMAN 2012 (2012); Accepted as regular paper (2012)
2. Thrun, S., Bennewitz, M., Burgard, W., Cremers, A.B., Dellaert, F., Fox, D., Hahnel, D., Rosenberg, C., Roy, N., Schulte, J., et al.: MINERVA: A second-generation museum tour-guide robot. In: Proceedings of the 1999 IEEE International Conference on Robotics and Automation, pp. 1999–2005. IEEE, Detroit (1999)
3. Shiomi, M., Kanda, T., Ishiguro, H., Hagita, N.: Interactive humanoid robots for a science museum. In: HRI 2006, pp. 305–312. ACM, Salt Lake City (2006)
4. Burgard, W., Cremers, A.B., Fox, D., Hähnel, D., Lakemeyer, G., Schulz, D., Steiner, W., Thrun, S.: Experiences with an interactive museum tour-guide robot. Artificial Intelligence 114, 3–55 (1999)
5. Mutlu, B., Forlizzi, J., Hodgins, J.: A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In: 2006 6th IEEERAS International Conference on Humanoid Robots, pp. 518–523. IEEE (2006)
6. Kuno, Y., Sadazuka, K., Kawashima, M., Yamazaki, K., Yamazaki, A., Kuzuoka, H.: Museum Guide Robot Based on Sociological Interaction Analysis. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2007, pp. 1191–1194. ACM, New York (2007)
7. Duffy, B.R.: Anthropomorphism and the social robot. Robotics and Autonomous Systems 42, 177–190 (2003)
8. Kanda, T., Miyashita, T., Osada, T.: Analysis of Humanoid Appearances in Human – Robot Interaction. IEEE Robotics 24, 725–735 (2008)
9. Uyen Tran, L., King, H.: The Professionalization of Museum Educators: The Case in Science Museums. Museum Management and Curatorship 22, 131–149 (2007)
10. Best, K.: Making museum tours better: understanding what a guided tour really is and what a tour guide really does. Museum Management and Curatorship 27, 35–52 (2012)
11. Wynn, J.R.: City Tour Guides: Urban Alchemists at Work. City & Community 9, 145–164 (2010)
12. Courage, C., Baxter, K.: Understanding Your Users: A Practical Guide to User Requirements Methods, Tools, and Techniques. Morgan Kaufmann (2005)

Getting Acquainted with a Developing Robot

Kerstin Fischer[1] and Joe Saunders[2]

[1] University of Southern Denmark, Sonderborg, Denmark
kerstin@sitkom.sdu.dk
[2] University of Hertfordshire, Hatfield, Hertfordshire, UK
j.1.saunders@herts.ac.uk

Abstract. Two factors that been suggested to influence the ways in which people interact with robots, namely users' initial expectations on the one hand and their increasing acquaintance with their robotic partner due to repeated interaction over time on the other. In the current study, eight participants interacted with a humanoid robot in five different sessions. Between the sessions, the robot was trained on the linguistic material presented to it by its human tutor in the preceding session, and thus the robot exhibits increasingly more knowledge of the domain. The results uncover the interaction between users' preconceptions and feedback-driven interactional effects that shape human-robot interactions. While considerable differences between users can be observed, all users respond to the robot's feedback and increasing linguistic capabilities in comparable ways.

1 Introduction and Previous Work

In this paper, we investigate how people interact with a developing robot. In order to study the role of increasing acquaintance, we analyze users' linguistic strategies by means of which they teach the robot over time. This will show us in how far the robot's behavior and increasing capabilities influence the way people interact with it and thus which impact social communication over time and, in particular, acquaintance with the robot may have.

Studies in cognitive psychology have shown that acquaintance plays a crucial role in the way in which people make use of common ground (see Clark [2]). Acquaintance has also been found to be a factor in studies of human-computer interaction; for instance, Amalberti et al. [1] compare participants' linguistic behaviors when they believe that their communication partner is either another human or a computer; they find that the considerable linguistic differences between speech directed at a computer and speech directed at another human, which can initially be observed in participants' speech, disappear gradually over several sessions. Thus, there is evidence that acquaintance plays a crucial role in interaction. However, it is so far unclear how such interactional effects are related to the preconceptions and expectations people bring into the interaction; several studies have shown that users' expectations also play a crucial role in the ways in which they interact with a communication partner (e.g. Fischer [6], Turkle [19]). This holds for interactions with

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 125–133, 2012.

communication partners with slightly different capabilities than one's own, such as foreigners (e.g. Zuengler [20]), as well as for interactions between younger and elderly people [11], but has also been shown for interactions with robots. For example, Turkle [19] argues that people's personal needs shape the ways they interact with relational artifacts, such as social robots. Fischer [6] shows that people's preconceptions about the degree of socialness of the human-robot interaction situation are an important factor in determining the way these people talk to a robot. Paepke and Takayama [13] manipulated users' expectations about the robot 'Pleo' by means of different introductory leaflets and find significantly different evaluations of the same robot after the interaction. Thus, preconceptions and users' expectations may have a considerable impact on HRI, yet it is unclear in how far these preconceptions are related to, and influenced by, what is happening in the course of the interactions between humans and robots.

The current study therefore aims to identify the effects of repeated interaction while taking people's initial expectations into account. We address this problem by investigating interactions between humans and a humanoid robot over time. In the current study, eight participants interacted with a humanoid robot in five different sessions. Between the sessions, the robot was trained on the linguistic material presented to it by its human tutor in the preceding session, and thus the robot exhibits increasingly more knowledge of the domain.

2 Data Elicitation

Eight adult participants took part in the study. Participants were between 27 and 58 years old (five female and three male). The backgrounds of the participants were either administrative (6) or research related (2), the latter not connected with robotic language research. Each of the eight participants took part in five interaction sessions of approximately two minutes with the robot (in total 40 robotic interaction sessions), and all of the sessions were videotaped for later analysis. The experiment was carried out over a three month period between March and June 2009 based on the availability of the participants. Participants were paid a small stipend of £20 if they completed all sessions (which all participants did).

In the experiment we asked the participants to teach the humanoid robot Kaspar (Dautenhahn et al. [4]) a series of shapes pasted on boxes. The robot was pre-programmed to track and habituate for a given period on these shapes. There was no constraint on participants' language. How to talk to the robot and what teaching strategies to use, was thus entirely up to the respective participant.

Following each interaction, the speech stream of the human was converted into phoneme strings marked with word boundaries. These phoneme strings were subsequently aligned with the sensorimotor modalities experienced by the robot during the interaction session. The aligned speech and sensory modalities were then processed to highlight words of long duration and words that appeared at the end of utterances. This processed modality stream became the basis for the robot's learnt experiences for the next interaction session with the human. In other words, the robot learned to associate the stressed words in a particular participant's speech stream with its visual perception of the shape presented to it during the sessions.

Fig. 1. A participant teaching Kaspar about shapes

In subsequent sessions (from session 2 onwards) the robot then matched its current sensorimotor input (that it was experiencing during the interaction) against that learnt in the previous session(s) with the particular tutor. This allowed the robot to react to the human by expressing (via its own speech) what it had learnt during the previous session(s). Thus, the robot produced feedback to the respective teacher by repeating words it had previously learned from associations of sounds to sensorimotor data. Full details of the experimental procedure can be found in Saunders et al. [16, 17].

3 Method

The method for analysis makes use of the principle of recipient design [15], which holds that people choose the linguistic features of their utterances to be suited best for their particular communication partners; for instance, people design their speech differently when speaking to children than when speaking to other adults (e.g. Snow [18]). In the current investigation, we make use of this principle by analyzing the participants' speech to the robot in order to identify who the participants think they are talking to. Thus, in the same way as we can identify speech to children by the shorter utterances, lower type-token ratio, lower complexity, more interactivity and more attention getting devices, we can study the properties of speech to a robot as a window into participants' concepts about their artificial communication partner and their ideas about what it will be good at and what it will have problems with. Thus, participants' linguistic choices reveal their concepts of their communication partner. The procedure thus consists in analyzing those linguistic features that may be revealing regarding participants' concepts of the robot and in identifying which of these features are affected by the variables investigated, here: the acquaintance with the robot. In some sense, this method is exploratory, as the main aim of the statistical analysis is to identify the nature of the adjustments participants make, rather than

testing specific hypotheses. On the other hand, the linguistic features analyzed have certain functions, and thus certain predictions can be made with respect to the areas in which changes take place.

4 Data Encoding

The data were orthographically transcribed and analyzed semi-automatically using shell scripts, whose results were manually controlled for correctness. The features investigated concern different linguistic features that may safely be assumed to be indicators of certain communicative functions and of people's conceptualizations and understandings of the robot and of the human-robot interaction situation. In particular, unambiguous linguistic features were automatically extracted from the transcripts if these are revealing with respect to participants' preconceptions and expectations about the robot, the task and the human-robot interaction. Since the linguistic features were extracted automatically, human contribution to this step is minimal, so that there is no manual encoding that would need to be checked by a second encoder. The only qualitative judgments made concern the selection of linguistic features investigated, which are therefore explained in detail below.

First, we looked for indicators that provide useful measures for the level of competence ascribed to the robot. These comprise structuring functions, for instance, items like *now, next*, but also *another*. These structuring cues presuppose that the interaction partner keeps track of the interaction and builds up a coherent representation of what he/she/it encounters. Another indicator of ascribed competence in the current scenario are ascriptions of memory and learning. For instance, if the robot is asked whether it memorizes something it had previously been told, this shows that participants expect that the robot learns and remembers what they teach it. Uses of past tense that refer to previous teaching sessions are indicators of such beliefs.

Second, in order to determine the social effects of the interaction, we investigated in how far users involve the robot directly. For instance, we counted instances of the personal pronoun *you*, instances of feedback signals, such as *good, well, excellent*, as well as instances of *yes* and *no*. Furthermore, we analyzed how often participants ask the robot questions, such as probing questions like *what's this?* and tag questions like *isn't it?*. Moreover, we looked at how often users call for the robot's attention by means of *look* or the robot's name.

We furthermore calculated the number of different words and, on the basis of the total number of words, the type-token ratio. The number of turns and the number of words are used to inform us on the one hand on how much effort the user put into the interaction, on the other, these numbers are used to calculate normalized numbers of the other features investigated, so that the numbers presented are always relative to the total number of turns or words used. The total numbers of turns and of the words used, as well as the type-token ratio, provide good indicators for how easy or difficult users make their utterances for their robotic partner. In speech to children, for instance, the number of different words and the type-token ratio are usually much lower than in speech to other adults (e.g. Snow [18]). Especially the diversity

measure, i.e. the type-token ratio, thus tells us whether users simplify their speech for the robot. These features thus function as indicators of suspected competence. They are common measures in readability tests, and speech adjusted to linguistically somewhat limited communication partners, such as children, is generally simplified in these terms. The same holds for the mean length of utterance (MLU), which is reliably reduced in speech to children (cf. Snow [18]; Roy et al. [14]).

We finally encoded whether participants greeted the robot at the beginning of each session. Whether a user greets a robot or not has been found to be a reliable indicator of the degree of socialness attributed to the robot, and as a useful predictor of the way this user will interact with the robot throughout the dialogs (Fischer [5, 6]; Lee et al. [12]).

5 Results

In order to assess the amount by means of which participants adjust their speech to the robot's behaviors over time, we compared the different sessions with each other, thus determining the likelihood that the interactions all stem from the same session. The results show that participants adjust their speech to the robot over time such that general tendencies in users' behaviors over time can be observed (see Table 1).

Table 1. Changes over time

	F(4,35)	p
turns	3.759235	0.012026
hello	0.261682	0.900508
words	1.150642	0.349159
diff_words	0.639448	0.637883
robot	1.000000	0.420651
now	0.770968	0.551458
another	1.607017	0.194327
interest	0.761221	0.557603
past	1.566045	0.204993
robot's name	0.764929	0.555259
look	3.204979	0.024169
lets	0.233275	0.917758
tag question	1.942775	0.125081
probing	1.822449	0.146530
expository	1.434195	0.243253
you	0.888735	0.480868
we	0.449707	0.771871
I	0.363485	0.832902
feedback	3.269179	0.022272
yes	1.406362	0.252151
no	0.891855	0.479093
MLU	5.429102	0.001651
typetoken	0.713872	0.588075

The analysis of the linguistic features shows that some significant changes occur over the five sessions. In particular, participants adjust the amounts of speaking such that the initial interactions are significantly shorter than especially the second interactions, and then interactions stabilize at a relatively high level. Thus, users spend different amounts of effort in the teaching sessions. Second, in the initial sessions, participants use significantly more devices by means of which they try to get the robot's attention; the number of instances of *look* is two-to-four times higher in the first session than in later sessions. In contrast, the number of feedback signals increases significantly over time, and most likely in correspondence to the robot's increasing linguistic capabilities. Finally, the mean length of utterance changes significantly after the first session and is adapted to the robot's linguistic capabilities in the later sessions.

As Table 1 shows, there are however no statistically significant differences in the amounts of structuring cues and references to the past, the use of the robot's name and other indicators of social relationship, tag questions and probing questions, pronouns, teaching strategies and linguistic diversity. Table 2 presents the means and standard deviations for the four features that change significantly during the five sessions.

So people adjust their speech according to the developing capabilities of the robot, in particular with respect to the amount of effort put into the interaction (number of turns), their perception of the need to keep the robot's attention, the amount of feedback given, and a central complexity measure, namely the mean length of utterances. At the same time, other linguistic features, which are generally subject to adjustments in child-directed speech, for instance, are not affected by the robot's increasing linguistic capabilities. Thus, participants do not structure the task more, do not reduce the number of different words, do not conceptualize themselves and the robot more as a team (as indicated by uses of 'let's' and 'we'), nor do they show differences in interpersonal relationships, such as by calling the robot's name, greeting it more, or referring less to themselves (by means of 'I') and more to the robot (by means of 'you'). While these features have been found to be affected by other aspects of robot behavior and embodiment, such as contingency of feedback and degrees of freedom (cf. Fischer, Lohan and Foth [9]; Fischer and Lohan [10]), they are obviously not affected by the robot's word learning.

Table 2. The four features 'number of turns', 'look', 'feedback' and 'MLU' across the five sessions

sessions	turns		look		feedback		MLU	
1	33.125	(5.16)	0.138	(0.14)	0.008	(0.016)	7.261	(1.528)
2	46.250	(8.28)	0.033	(0.03)	0.013	(0.020)	4.786	(1.137)
3	42.375	(8.57)	0.076	(0.08)	0.042	(0.063)	5.235	(1.393)
4	44.750	(5.39)	0.031	(0.03)	0.058	(0.051)	4.773	(1.409)
5	43.750	(9.41)	0.021	(0.04)	0.082	(0.066)	4.295	(1.546)

However, besides for functional reasons, the failure to find more statistically significant differences between sessions may be due to high interpersonal variation. In a next step, we therefore investigated interpersonal differences in the interactions.

In order to assess the interpersonal differences between the eight different participants, we compared their linguistic behaviors in the five sessions with each other.

The investigation of differences in the linguistic features between participants shows that there are considerable differences between users throughout. In fact, only tag questions, number of turns, instances of 'look' and instances of '*yes*' are not significantly different between participants.

Thus, the analysis shows extreme interpersonal differences between speakers, basically concerning all linguistic choices. This suggests that participants differ considerably in their understanding of the situation (cf. Fischer [6]). However, while people differ in almost all linguistic behaviors, with respect to two of the four features that were found to be adjusted to the robot over time people converge in their linguistic choices; in fact, we can also understand the lack of differences in the use of 'yes' from the same perspective since the most important function of 'yes' is to provide feedback. The robot's developing capabilities can consequently be taken to guide people subtly into similar behaviors.

Table 3. Interpersonal Differences

	F(7,32)	p
turns	0.70765	0.665664
hello	2.92517	0.017482
words	4.47183	0.001450
diff_words	12.30139	0.000000
now	4.76811	0.000929
another	4.20360	0.002190
interest	2.56840	0.032148
past	3.97814	0.003117
robot's name	3.12951	0.012395
look	1.95766	0.092588
lets	4.37695	0.001676
tag questions	1.00065	0.448921
checking	2.54773	0.033312
expository	3.03480	0.014529
you	3.92834	0.003372
we	16.42579	0.000000
I	4.53277	0.001322
feedback	2.83286	0.020446
yes	1.56765	0.180879
no	6.10001	0.000142
MLU	2.62818	0.029008
typetoken	11.17388	0.000000

6 Discussion

The linguistic analyses presented show that the human tutors adjust their instructions to the robot's linguistic behavior over time. The linguistic features changed are functionally related to the different communicative tasks that users encountered in the five sessions. In particular, in the first session, users' communicative efforts largely concerned getting the robot's attention, which corresponds to the fact that the robot's only means of feedback was to display its attention nonverbally. So users' communicative focus in the first session is consistent with users' orientation at the robot's behavior (Fischer et al. [8]). These communicative efforts change already in the second session when the robot starts producing verbal output.

The other changes made by the participants over the course of the sessions concern the mean length of utterance, the amount of speaking and the amount of linguistic feedback. These changes can be related to different tutoring behaviors on the one hand and the robot's increasing linguistic capabilities on the other. The changes observed are thus in accordance with a model of human-robot interaction that assumes high amounts of cooperation from the side of the users (cf. Fischer [5]) and considerable attention to the robot's capabilities (Fischer [7]; Fischer et al. [8]).

The results concerning interpersonal variation have shown that users' expectations and preconceptions play a considerable role in interaction. However, irrespective of their different preconceptions, all users converge on the same behaviors in response to the robot's behavior.

7 Conclusion and Future Work

We can conclude that both users' preconceptions and feedback-driven interactional effects shape human-robot interactions. While the initial differences between users persist over time, all users respond to the robot's feedback and increasing linguistic capabilities in comparable ways. Thus, the good news for robot developers is that the kinds of behaviors the robot produces subtly guide users into similar kinds of responses, irrespective of their initial expectations. Future work will have to identify the factors that lead to the high interpersonal variation identified – what makes participants understand the same human-robot interaction situation so differently that they make significantly different linguistic choices for their partners that persist over time? Furthermore, besides understanding interpersonal variation, it will also be useful if people's differing behaviors can be predicted; on the other hand, the current results suggest that the robot's behavior can guide people into particular behaviors; future work should thus explore in more depth how participants' ideas of the HRI situation and the robot's capabilities can be shaped.

References

1. Amalberti, R., Carbonell, N., Falzon, P.: User Representations of Computer Systems in Human-Computer Speech Interaction. International Journal of Man-Machine Studies 38, 547–566 (1993)

2. Clark, H.H.: Arenas of Language Use. Cambridge University Press (1992)
3. Clark, H.H.: Using Language. Cambridge University Press (1996)
4. Dautenhahn, K., Nehaniv, C.L., Walters, M.L., Robins, B., Kose-Bagci, H., Mirza, N.A., et al.: Kaspar - a minimally expressive humanoid robot for human-robot interaction research. Applied Bionics and Biomechanics, Special Issue on 'Humanoid Robots' 6(3), 369–397 (2009)
5. Fischer, K.: What Computer Talk is and Isn't: Human-Computer Conversation as Intercultural Communication. AQ, Saarbrücken (2006)
6. Fischer, K.: Interpersonal Variation in Understanding Robots as Social Actors. In: HRI 2011, Lausanne, Switzerland (2011a)
7. Fischer, K.: How people talk with robots – Designing Dialog to Reduce User Uncertainty. AI Magazine 32(4), 31–38 (2011b)
8. Fischer, K., Foth, K., Rohlfing, K., Wrede, B.: Mindful tutors - linguistic choice and action demonstration in speech to infants and to a simulated robot. Interaction Studies 12(1), 134–161 (2011)
9. Fischer, K., Lohan, K., Foth, K.: Levels of Embodiment: Linguistic Analyses of Factors Influencing HRI. In: HRI 2012, Boston (2012)
10. Fischer, K., Lohan, K.: How Robot Embodiment and Situatedness Influence Interaction. In: Cognitive Linguistics Yearbook. Mouton de Gruyter, Berlin (submitted)
11. Giles, H., Coupland, J., Coupland, N.: Contexts of Accomodation. Developments in Applied Sociolinguistics. Cambridge University Press, Cambridge (1991)
12. Lee, M.K., Kiesler, S., Forlizzi, J.: Receptionist or Information Kiosk: How Do People Talk with a Robot? In: CSCW 2010, Savannah, Georgia, February 6-10 (2010)
13. Paepcke, S., Takayama, L.: Judging a Bot By Its Cover: An Experiment on Expectation Setting for Personal Robots. In: Proc. of Human Robot Interaction (HRI), Osaka, Japan (2010)
14. Roy, B., Frank, M., Roy, D.: Exploring word learning in a high-density longitudinal corpus. In: Proceedings of the 31st Annual Meeting of the Cognitive Science Society, Amsterdam, Netherlands (2009)
15. Saunders, J., Lehman, H., Sato, Y., Nehaniv, C.: Towards Using Prosody to Scaffold Lexical Meaning in Robots. In: Proceedings of ICDL-EpiRob 2011: IEEE Conference on Development and Learning, and Epigenetic Robotics (2011)
16. Sacks, H., Schegloff, E.A., Jefferson, G.: A Simplest Systematics for the Organization of Turn Taking for Conversation. Language 50, 696–735 (1974)
17. Saunders, J., Nehaniv, C.L., Lyon, C.: Robot learning of lexical semantics from sensorimotor interaction and the unrestricted speech of human tutors. In: Proc. Second International Symposium on New Frontiers in Human-Robot Interaction, AISB Convention, Leicester, UK (2010)
18. Snow, C.E.: Beginning from baby talk: Twenty years of research on input and interaction. In: Gallaway, C., Richards, B.J. (eds.) Input and Interaction in Language Acquisition, pp. 3–12. Cambridge University Press, Cambridge (1994)
19. Turkle, S. (2006): A Nascent Robotics Culture: New Complicities for Companionship. AAAI Technical Report series (July 2006)
20. Zuengler, J.: Accomodation in Native-Nonnative Interactions: Going beyond the "What" to the "Why" in Second-Language Research. In: Giles, H., Coupland, J., Coupland, N. (eds.) Contexts of Accommodation. Developments in Applied Sociolinguistics. Cambridge University Press, Cambridge (1991)

Learning the Combinatorial Structure of Demonstrated Behaviors with Inverse Feedback Control

Olivier Mangin[1,2] and Pierre-Yves Oudeyer[1]

[1] Flowers Team, INRIA, France
[2] Université Bordeaux 1, France

Abstract. In many applications, such as virtual agents or humanoid robots, it is difficult to represent complex human behaviors and the full range of skills necessary to achieve them. Real life human behaviors are often the combination of several parts and never reproduced in the exact same way. In this work we introduce a new algorithm that is able to learn behaviors by assuming that the observed complex motions can be represented in a smaller dictionary of concurrent tasks. We present an optimization formalism and show how we can learn simultaneously the dictionary and the mixture coefficients that represent each demonstration. We present results on a idealized model where a set of potential functions represents human objectives or preferences for achieving a task.

1 Introduction

Robots are expected to have promising applications in fields such as domestic assistance, health care or education. However bringing robots to our everyday environment and improving their interaction capabilities requires that they are capable of understanding natural human behaviors.

Human activities are numerous and highly diverse, and feature large variability between individuals, situations, and times. Making robots or intelligent systems capable to recognize or even understand or reproduce such behaviors, thus requires a high level of adaptivity which makes learning algorithms promising candidates for this task.

It is however still a difficult problem to design or adapt learning algorithms so that they can deal well with essential properties of natural human behaviors. In fact natural human behaviors are complex and one won't generally observe something as "fill a glass of water" but rather "grasp a glass, walk to the tap, open the tap while keeping the glass straight". Being able to cope with the combinatorial structure of behaviors is thus necessary for their understanding.

In both examples each primitive behavior must be separated from the other behaviors composing the general activity and the relevant features must be identified as the glass being filled, not as the exact trajectory of the elbow or the position of the glass. These two difficulties are actually related to wider topics of research from which efficient algorithms and representations can benefit

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 134–147, 2012.

human behavior understanding by leveraging compositional structure of human activities and represent tasks or objectives that drive the activities.

First separating complex behaviors into simpler parts is very close to both the decomposition of complex motions into simpler motor primitives and dictionary learning techniques from machine learning.

Then, focusing on representations of behaviors in terms of the cost function they are optimizing rather than the specific way to solve it is closely related to inverse feedback control and inverse reinforcement learning approaches which can lead to better generalization properties, as for example when learning to imitate.

In this article we address aspects of the issues of representing, learning and reproducing human behaviors and their compositional structure. We introduce a dictionary learning approach for representing and reproducing the combinatorial structure of motor behaviors that are only observed through demonstrations of several concurrent motor behaviors. We focus on motor behavior representations that directly model the objective of the user underlying demonstrations. We illustrate the presented algorithm on a simple toy example.

2 Background and Related Work

2.1 Decomposition of Motor Skills: Motor Primitives

Motor primitives have been introduced as a form of re-usable motor skills that may be used as elementary building blocks for more complex motor control and skills. The concept of motor primitives that can be combined together has the appealing property to enable combinatorial growth of the skill repertoire. As detailed by Konczak [1], examples of motor primitives can be found both in biological and robotic systems, and can be either innate or acquired.

The notion of combination of motor primitives can take different forms. One could consider a behavior composed of a **sequence** of simple actions, like moving one's hand to a glass, grasping it, bringing it back to one's mouth, etc.

The structure of some behaviors however does not fit well in this sequential representation. Many behaviors or tasks are better described in terms of elementary movements executed **simultaneously** (e.g. on different parts of the body) or **concurrently**, like speaking while smiling and shaking someone's hand. Concurrent combinations of behaviors is particularly studied in this article.

2.2 Using HMMs to Learn Motor Primitives

Hidden Markov models (HMM), often coupled with clustering techniques or mixture models, have been largely used to learn sequences of primitives. For example, Kulic and Nakamura have proposed in [2] a method that first performs an unsupervised segmentation of the motion signal into small successive blocks (the segmentation technique itself is based on HMMs), and then performs clustering over HMM representations of each segmented block. Each group of similar motions is interpreted as a motor primitive.

In a different setting, Kruger et al. [3], have focused on a notion of motor primitive based on the effect of actions on objects from the environment. They have proposed to first discover primitives by clustering action effects on manipulated objects and then use the found clusters, composed of actions that have similar effects, to segment the stream of motions into coherent actions. Then parametrized hidden Markov models are trained to represent the actions and enable both their recognition and reproduction.

Finally Calinon et al. [4] and Butterfield et al. [5] use Gaussian mixture models to represent motion primitives and HMMs to discover and represent the transitions and sequential combinations of primitives. All the approaches presented in this paragraph are capable of recognizing and reproducing the learned motions.

2.3 Using Dictionary Learning to Learn Motor Primitives

Dictionary learning approaches by matrix factorization are machine learning techniques widely used to solve problems where an input signal has to be decomposed into a linear combination of atoms. They target the learning of both the dictionary of atoms and the coefficients used in their combinations.

The possibility to enforce structural constraints on the dictionary and coefficient matrices enables better modeling of many problems and participates in the versatility of dictionary learning techniques. Such constraints include for example non-negativity [6,7], sparsity or group sparsity [8], constraining atoms to be convex combinations of the demonstrations, which can be seen as a generalization of clustering [9], constraining atoms to be stochastic vectors, etc.

In the field of motion decomposition, Li et al. [10] have used orthogonal matching pursuit to decompose complex motions into simple motion patterns activated shortly along time. The decomposition is used to perform both compression, classification and reproduction of visualizations of the movement (but is not tested on real reproduction). The article uses constraints such as sparse activation coefficients and sparse motion patterns in Fourier domain.

Hellbach et al. [11] have also used non-negative matrix factorization to perform a decomposition of globally unstructured motions in low level components. They use time invariance and sparsity of dictionary atoms to guide the learning toward discovering short sequences of positions that can be concatenated into the observed trajectory. These capabilities are tested on a dataset of real movements for prediction but not to produce motion on a real robot.

Time sequences of motor primitives learnt by methods from Li et al. [10] and Hellbach et al. [11] may include overlap, and can therefore be considered as hybrid methods enabling the learning of motor primitives combined both in sequence and parallel. They are however mainly focused on representing trajectories by superposition of time shifted simple local patterns and do not explore how the structure of complex behaviors composed of simultaneous primitive gestures can be leveraged towards better understanding of the observed activity.

In our previous work [12] we demonstrated how non-negative matrix factorization can be used to decompose complex behaviors into simultaneous combinations of primitive gestures. We presented an experiment in which dance

choreographies are demonstrated by a human. Each choreography is composed of several simultaneous gestures. For example, one leg gesture and one gesture on each arm. A set of symbolic linguistic labels corresponding to the gestures occurring in the choreography are also provided with the demonstrations, which is a form of linguistic guidance. A learning system is trained by observing both demonstrations of the choreography and the associated labels. The system then observes new dances and has to reconstruct the associated set of labels, that is to say, tell which gestures were combined to form the choreography. It is shown in the article that the system performs well even if the demonstrated choreography is a combination of gestures that have never been demonstrated together during training. This setting emphasizes the ability of the system to capture the compositional structure of the choreographies.

[12] presents a technique that permits classification of complex behaviors, but it cannot reproduce them since the motion representation is only discriminative. This article presents a dictionary learning approach based on inverse feedback control, which as a generative representation enables motion reproduction.

2.4 Inverse Feedback Control

Approaches which consist in direct representation and reproduction of the policy (state to action mapping) observed through trajectories of the demonstrator's (or imitator's) body are often denoted as **policy learning**. Most techniques presented in Sections 2.2 and 2.3 belongs to this category. The policy can either be a direct representation of the trajectory [13] or a probabilistic model of the policy [4].

In opposition, **inverse optimal control** [14] and **inverse reinforcement learning** [15] are approaches based on the idea that, in some situations, it can lead to better generalization to model aspects of the task that the demonstrator is trying to solve instead of modeling the particular solution in the demonstrated context. The capacity of inverse optimal control to achieve better generalization has been demonstrated in the experiment performed by Abbeel et al. [16], in which an helicopter performs acrobatic motions after observing demonstrations from a human expert remotely controlling the helicopter. In that example the learned trajectories even overtake the skills of the demonstrating expert.

Jetchev and Toussaint [17] have adapted inverse optimal control techniques to a single grasping task on a real robot. Furthermore they have shown how the inverse optimal control approach, coupled with a sparsity constraint on the task representation can be used to discover relevant features in the task space.

Finally Brillinger [18] has developed an algorithm based on least square regression to learn potential functions modeling the motion of wild animals in natural parks.

In this article we extend Brillinger's technique to address a different problem: instead of learning a flat representation of a single task, the learner must infer several primitives cost functions/skills that can be composed to explain the mixing of concurrent tasks that are demonstrated. We use a very similar behavior representation, but introduce dictionary learning for solving the new problem.

3 Problem Definition and Algorithm

We introduce a simple synthetic imitation learning experiment in which an imitator learns to reproduce behaviors observed from a demonstrator.

More precisely we model the task underlying each behavior as a cost function on states of the agent (either the demonstrator or the imitator), which can be seen as representing the preferences of the demonstrator. For example the task of filling a glass of water will be represented by a cost function giving increasing values to increasing levels of water in the glass. In the case where the "filling the glass" behavior is mixed with the "smiling to someone" behavior, the mixed behavior will be represented by a mixed cost function valuing both full glass and smiling position of the lips.

Each demonstration consists in a trajectory in the demonstrator state space, from a specific initial position. The objective of the imitator is to produce a trajectory (either from the same initial position than the demonstration, or another) that fits the demonstrator preferences (i.e. minimize the cost function).

This setup introduces two important difficulties for the imitator. On the one hand each demonstration only presents aspects of the cost function locally, around the trajectory. Each demonstration is thus not sufficient to fully understand the underlying task. On the other hand, each demonstration presents a mixture of several tasks. Thus, while the primitive tasks are observed many times, they are never observed alone and each particular mixture is generally only observed once. It is thus necessary to leverage the compositional structure of the behaviors to be able to understand them, and reproduce them with new initial positions.

3.1 Agent and Demonstrator Models

We will assume that both the demonstrator and imitator are identical. This corresponds for example to the case where demonstrations are performed on the imitator body (kinesthetic demonstrations). Following Jetchev et al. [17], we consider a robotic agent which configurations q belong to a state space $\mathcal{Q} \in \mathbb{R}^S$. Each trajectory is denoted by a sequence $(q_t)_{t \in [|1,T|]}$.

We assume that there exists a cost function $f : \mathcal{Q} \longrightarrow \mathbb{R}$ such that each task is modeled as the demonstrating agent trying to minimize the cost $f(q)$ to which is added a penalization on the square norm of $\frac{\partial q}{\partial t}$, which can be seen as a penalization of the energy consumed while moving to optimize $f(q)$.

We will focus on very simple agents which actions are motions in the state space and are governed by the local optimization of $f(q) + \alpha \left\| \frac{\partial q}{\partial t} \right\|^2$ which means that each action, at each time step, is chosen such that:

$$q_{t+1} = \underset{q}{\operatorname{argmin}}\ f(q) + \alpha \left\| \frac{q - q_t}{\delta_t} \right\|^2,$$

where δ_t is the time elapsed between samples t and $t+1$.

The solution of this equation, without additional constraints, and assuming that the cost function f is differentiable, is well known to be proportional to the gradient of f, as $-\frac{1}{\alpha}\nabla f(q)$.

It can be noticed that since the agent we have defined only follows policies driven by local optimization it will only achieve local optimization of the cost function. While this is a simplification of the agent, it also features an important property of real demonstrators: real demonstrators are in general imperfect and do not always succeed in reaching the optimal solution of the task. It is thus important for a imitator to be able to also learn from imperfect demonstrations of behaviors.

In this article we focus on complex tasks: each demonstration corresponds to the minimization of a separate cost function f which is only observed through one demonstration. However f is composed of parts that also occur in other demonstrations and are thus observed several time mixed in various way and in various contexts.

Lets consider N demonstrations, observed as trajectories $(q_t^i)_t,\ i \in [|1, N|]$ in the agent state space. We assume that each demonstration corresponds to a given f^i. To model complex demonstrations we assume that there exists a dictionary of primitive tasks, composed of K cost functions $(g^k)_{k\in[|1,K|]}$, such that, for all demonstration i, there exist coefficients $(a_k^i)_{k\in[|1,K|]}$ such that, for all state q, $f^i(q) = \sum_{k=1}^{K} a_k^i g^k(q)$.

We present a learning algorithm which observes one demonstration associated with each function f^i and learns a dictionary of primitive cost functions g^k, and the coefficients of their combinations into demonstrated tasks f^i.

3.2 Inferring a Task from a Demonstration

The problem of inferring a single task from a demonstration is studied in Brillinger's article [18]. The cost function is represented by a linear parameter $\beta \in \mathbb{R}^F$ on a space of potentially non-linear features $\varphi : \mathcal{Q} \longrightarrow \mathbb{R}^F$. Its minimization is modeled by an agent policy such that:

$$\frac{\partial q}{\partial t} = -\lambda \mathbf{J}(q)^T \beta \tag{1}$$

where $\mathbf{J}$ is the Jacobian of φ (lines of $\mathbf{J}$ are gradients of coordinates of φ).

When discrete trajectories are considered, equation (1) approximates into $\frac{q_{t+1} - q_t}{\delta_t} = -\lambda \mathbf{J}(q_t)^T \beta$ for all $t \in [|1, T-1|]$. By denoting $y_{t+1} = \frac{q_{t+1}-q_t}{\delta_t}$, $Y \in \mathcal{R}^{S\times(T-1)}$ the vector obtained by vertically stacking all y_t for $t \in [|2, T|]$, and $\mathbf{\Phi}$ the $S \times (T-1)$ by F matrix obtained by vertically stacking all $-\lambda \mathbf{J}(q_t)^T$, we get:

$$Y = \mathbf{\Phi}\beta \tag{2}$$

Equation (2) transforms the problem of inferring one task from one demonstration into a linear regression problem, which constitutes an essential contribution of Brillinger's article.

In the case where the Euclidean distance between the vector Y, computed from observations, and its reconstruction through the task model $\mathbf{\Phi}\beta$ is considered, we get the classical least square regression problem. It is solved, assuming $\mathbf{\Phi}^T\mathbf{\Phi}$ is non-singular, by:

$$\beta = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T Y \tag{3}$$

More details on the associated derivations can be found in [18]. The algorithm presented above is capable, from one demonstration, to infer the cost function modeling a behavior of the demonstrator. Once the cost function is inferred, the imitator can in turn produce trajectories that minimize it. Such an agent that directly infers all the parameters of the cost function is denoted **flat imitator** in the following.

3.3 Learning a Dictionary of Primitive Tasks from Mixed Demonstrations

The algorithm presented in Section 3.2 only applies to a single demonstration generated from a single task model. In this section we introduce a matrix factorization algorithm to learn a dictionary of primitive tasks and associated coefficients from several demonstrations.

Each demonstration corresponds to a mixing of primitive tasks which is modeled by a β^i in the feature space. To model the concurrent mixing of primitive tasks, we introduce a dictionary represented by a F by K matrix $\mathbf{D}$ such that each column of $\mathbf{D}$ is the parameter representing the primitive tasks g^k in the feature space. The concurrency between the primitive tasks in a mixing is represented through a weighting coefficient. Coefficients of the i^{th} demonstrated task are given by a vector $a^i \in \mathbb{R}^K$, $\beta^i = \mathbf{D}a^i$.

For each demonstration we define the vector Y^i and the matrix $\mathbf{\Phi}^i$ associated with the observed trajectory, by following the method described in Section 3.2. It follows that for each demonstration:

$$Y^i = \mathbf{\Phi}^i\mathbf{D}a^i \tag{4}$$

Learning a factored model of the demonstrated tasks that minimize Euclidean distance to demonstration is equivalent to solving equation (5).

$$\underset{\mathbf{D},\mathbf{A}}{\operatorname{argmin}}\ \mathcal{L}(\mathbf{D},\mathbf{A}) \text{ with } \mathcal{L}(\mathbf{D},a) = \sum_{i=1}^{N} \left\| Y^i - \mathbf{\Phi}^i\mathbf{D}a^i \right\|_2^2 \tag{5}$$

We propose an algorithm based on alternate minimization with respect to $\mathbf{D}$ and $\mathbf{A}$ to solve this problem.

Minimization with respect to $\mathbf{A}$ This sub-problem assumes that the dictionary is known and thus consist, from a demonstration, in inferring the task decomposition on the dictionary. It is similar to the algorithm presented in previous section, but the K decomposition coefficients (the vector a) are inferred instead of all the F coefficients of the cost function.

This problem is separable in one sub-problem for each demonstration i which are all equivalent to the regression problem presented in Section 3.2 where the matrix $\mathbf{\Phi}$ is now replaced by the product $\mathbf{\Phi}^i\mathbf{D}$. Thus the solution of the optimization with respect to $\mathbf{A}$ is given, for Euclidean distance, by equation (6). Other norms or penalization could as well be used to solve the regression (e.g. methods enforcing non-negativity or sparseness of coefficients).

$$a^i = (\mathbf{D}^T{\mathbf{\Phi}^i}^T\mathbf{\Phi}^i\mathbf{D})^{-1}\mathbf{D}^T{\mathbf{\Phi}^i}^T Y^i \tag{6}$$

Minimization with respect to $\mathbf{D}$ The second sub-problem assumes that the decomposition coefficients of the demonstrated task are known but not the dictionary $\mathbf{D}$. We use a gradient descent approach to learn $\mathbf{D}$. The differential of the loss with respect to each of the coefficients of $\mathbf{D}$ is given by equation (7).

$$\nabla_{\mathbf{D}}\mathcal{L}(\mathbf{D},\mathbf{A}) = -2\sum_{i=1}^{N}{\mathbf{\Phi}^i}^T\left[Y^i - \mathbf{\Phi}^i\mathbf{D}a^i\right]{a^i}^T \tag{7}$$

Global algorithm The global algorithm simultaneously learns the dictionary $\mathbf{D}$ and the coefficients $\mathbf{A}$ by alternation of the two procedures from previous paragraphs. Matrices $\mathbf{D}$ and $\mathbf{A}$ are initiated randomly or according to any heuristic. Then $\mathbf{D}$ is learnt, assuming $\mathbf{A}$ contains the correct decomposition coefficients, after which $\mathbf{A}$ is inferred assuming $\mathbf{D}$ is the correct dictionary, and so on. This approach to matrix factorization problems has often proved to be efficient ([7,8]).

4 Experiments

To illustrate the algorithm introduced in Section 3 we consider a simple toy experiment. We define an agent which state q belongs to $\mathcal{Q} = [0,1]^2$. Cost functions are parametrized on a 5 by 5 grid of Gaussian radial basis functions , which means $\phi(q)^T = (\dots, \frac{1}{2\pi\sigma}\exp(-\frac{\|x-\mu_f\|^2}{2\sigma^2}), \dots)$ where μ_f are points from a regular 5 by 5 grid on $\mathcal{Q}$ and σ is fixed such that the task parameter space is of dimension $F = 25$.

We use in this experiment a dictionary of 6 primitive tasks that is represented in Figure 1 (first row). Combinations of 2 or 3 concurrent primitive tasks are generated randomly for training and testing. For a given mixed tasks, a starting point is randomly chosen inside $\mathcal{Q}$ and trajectories are generated by the demonstrator or imitator from the initial position, according to equation (1). In the remaining of this section we will describe two separate experiments where a dictionary is learnt by a agent observing mixed combinations of tasks.

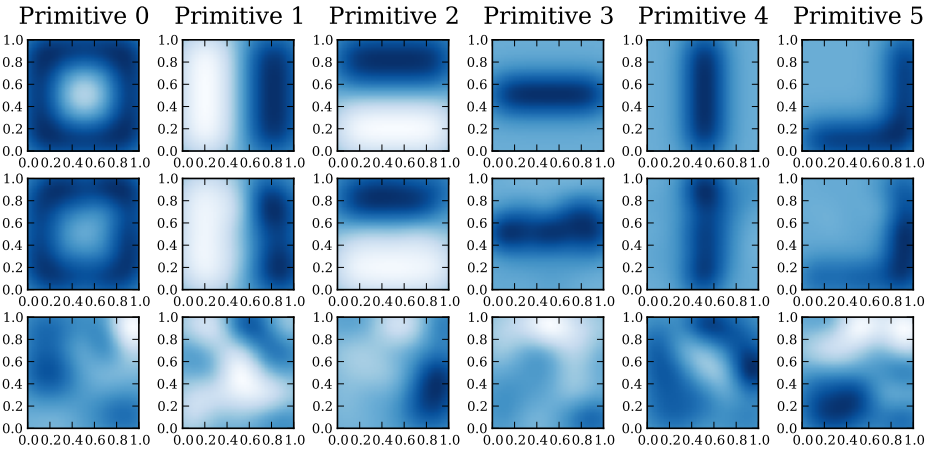


Fig. 1. Dictionary of primitive tasks represented as cost functions over $\mathcal{Q} = [0, 1]^2$. First row corresponds to original primitive tasks (as used by the demonstrator), second row to the one reconstructed by the learner described in Section 4.1 and third row to the learner described in Section 4.2. Dark areas correspond to high positive costs and light areas to negative costs. (Best viewed in color).

4.1 Recovering the Dictionary from Given Coefficients

In this section we consider an experiment in which during training the learner both observes demonstrations of mixed tasks and the associated mixing coefficients. This hypothesis models the situation where some labels associated with the task that are mixed together in the demonstration are given to the learner (e.g. inferred from spoken language). This experiment enables the evaluation of the second part of the algorithm we introduced.

Since the mixing coefficients are known by the learner during training, only the second part of the algorithm presented in Section 3.3 is used to learn the dictionary $\hat{\mathbf{D}}$. We train such a learner on 200 trajectories generated from a dictionary $\mathbf{D}$. Both the original dictionary of primitive tasks $\mathbf{D}$ and its reconstruction $\hat{\mathbf{D}}$ are represented in Figure 1.

Once the imitator has built a dictionary of tasks from observations, it is evaluated in the following way: for a set of coefficients, corresponding to mixed tasks, and a random starting position, the imitator and demonstrator yield trajectories. The demonstrator and imitator trajectories are then compared. Examples of trajectories from both the learner and the imitator are given in figure 2.

The relative L_2 error between the trajectories generated by the demonstrator and the imitator is used to evaluate the quality of the reconstruction. An average error of 0.001127 is obtained on the train set (tasks observed while learning the dictionary) and 0.002675 is obtained on the test set (unobserved tasks).

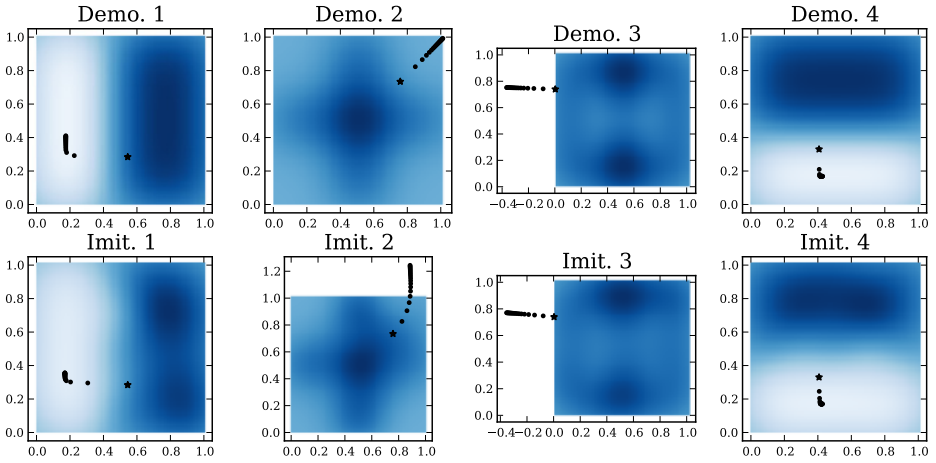


Fig. 2. Examples of demonstration trajectories generated from mixed concurrent primitives tasks (first row) and their reproduction by the learner from experiment one. Initial positions are marked by stars, others position are marked by circles. The associated cost functions (the one inferred in the case of the imitator) are also represented. Dark areas correspond to high positive costs and light areas to negative costs. (Best viewed in color).

4.2 Learning Both Primitive Tasks and Mixing Coefficients from Concurrent Demonstrations

We illustrate the full algorithm presented in Section 3.3 on an experiment where the learner only observes demonstrated trajectories without knowing the coefficients. The learner's reconstructed dictionary is given in Figure 1, bottom row.

Once the dictionary has been learnt, we use the following imitation protocol to test the imitator. A new unobserved combination of primitive tasks is chosen together with an initial position. Then the demonstrator provides a trajectory corresponding to the task. From the observation of the demonstrated trajectory and the learnt dictionary of primitive tasks, the learner infers the task's decomposition on the learnt dictionary (using the first part of the algorithm presented in Section 3.3). Finally the imitator is asked to produce trajectories corresponding to the same task, both from the demonstrator's initial position and randomly chosen initial positions. Changing the initial position from the demonstrated one is a way to evaluate how well the imitator's model of the task generalizes from the demonstration context to new ones.

In order to evaluate the impact of learning the dictionary, that is to say the combinatorial structure of the demonstrated data, we compare reproductions of the task by an agent that has learnt the dictionary denoted as *full dictionary learner*, to ones by an agent, denoted as *flat imitator*, that directly infers the parameters of the tasks without using a dictionary (algorithm presented in Section 3.2). We also compare the agent described in the previous section that

has learnt the dictionary from both demonstrated trajectories and mixed coefficients, denoted *dictionary from coefficients learner*. Examples of demonstrated and imitated trajectories are provided in Figure 3.

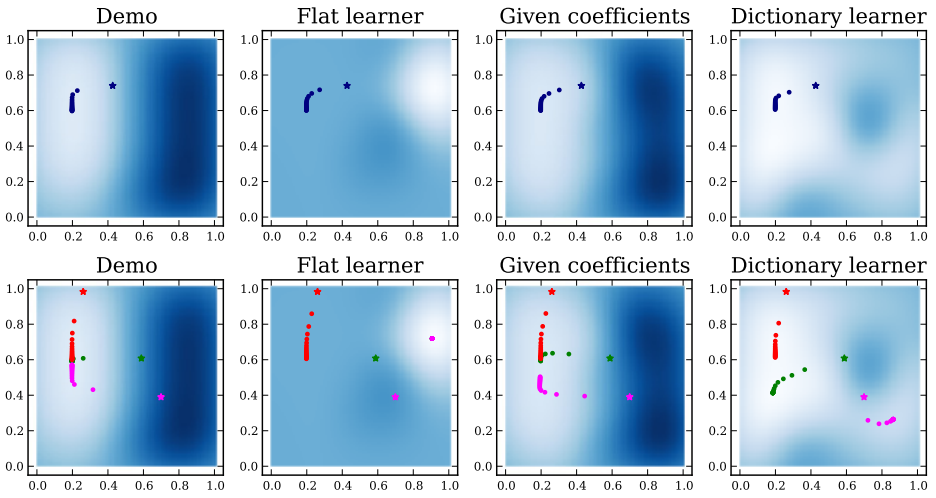


Fig. 3. Examples of imitated trajectories. First row presents the demonstrated trajectory (first column) and its imitation by the flat learner, the dictionary learner from first experiment (coefficients observed while learning the dictionary) and the full dictionary learner. Second row correspond to imitations of the same task from initial positions that were not observed (the demonstrator trajectories for those positions are given for comparison purpose). (Best viewed in color).

5 Discussion

The first agent presented in Section 4.1, is able, by observing motions solving composed tasks and the mixing coefficients, to learn the dictionary of primitive tasks. The acquired dictionary is evaluated in different ways: visually from the plots of the associated cost functions, from trajectories solving a mixed task whose mixing coefficients are given, and from imitation, in random contexts, of a mixed task that is inferred from a single demonstration (this last result is presented together with second experiment).

In our previous work [12], we present an algorithm that learns from mixed behaviors presented together with labels similar to the mixing coefficients. The learner is able to yield the labels from test demonstrations of the motions. Actually the experiment evaluates the algorithm directly on the quality of the estimation of the coefficients, since the system is not able to reproduce the demonstrated gestures. The first agent presented in this article learns in a similar setting than the algorithm from [12] but extends its capabilities to the reproduction of the demonstrated behaviors.

The second agent described in Section 4.2 is capable of learning a dictionary that enables the factorial representation of demonstrated tasks, without directly observing the dictionary or the mixing coefficients. The factorial representation enables imitation of tasks that are observed through a single demonstration. However the performance of the imitator is not evaluated due to the illustrative nature of the experimental setup. In particular the least square regression from [18] (described in Section 3.2) is not performing well on the particular form of mixing of cost functions we have chosen for the illustrative toy example. However our algorithm is compatible with any regression method. Thus, interesting further work could use the comparison of performances between various regression methods, on real human data, to get better insight on the combinatorial properties of human activities.

The dictionary learnt by the agent is very different from the one of the demonstrator. Actually the problem of representing a set of demonstrated mixed tasks as linear combinations of primitive tasks is ill posed and does not have a unique solution. For example one can scale the primitive cost function by some factor and associated coefficients by its inverse or change the order of the primitive and coefficients without changing the linear combination. Mathematically these difficulties could be solved by adding constraints to the form of the learnt dictionary (e.g. normalize primitive costs) or by adapting the way to compare dictionaries (e.g. to make it invariant to re-ordering).

To overcome this difficulty, several ways of making some possible decompositions more salient than others can guide the learning, in the same way humans easily identify salient behaviors even when mixed with others. First, saliency can come from one's history: if one already knows all but one primitive behavior present in the scene, it is possible to identify the unexplained parts of the behavior and learn it as a new primitive. Investigating this part would require to extend the learning model to an incremental learner. The algorithm we presented can be extend to become online following a similar method than [19] although this is not investigated in this article.

Then, a particular form of factorization could also be shaped by information coming from another modality or social interaction. This aspect is demonstrated both in our previous work [12] and in the first experiment (Section 4.1), where observing the mixing coefficients, that can be seen as linguistic labels, enables the learner to adapt its internal model (i.e. the dictionary) to a communication channel. Aspects of social learning have already been shown to improve motor learning by Massera et al. [20]. Solving the ambiguity in the decomposition of human activities thus constitutes a new application for social learning.

Finally intrinsic constraints can be applied to the learnt dictionary to prefer some solutions . Two examples of such constraints for which many machine learning algorithms have been developed are non-negativity and sparsity. Non-negativity of the coefficients will for example focus on representations that allow primitive behaviors to be added to but not subtracted from an activity in which they do not appear. Jetchev et al. [17] have shown how enforcing sparsity of a task representation can make this task focus only on a few salient features, thus

performing task space inference. Other examples are given by Li et al. [10] and Hellbach et al. [11].

Extending the algorithm we presented to include constraints or evaluating it on an online learning experiment would help investigating these questions and thus constitute very interesting future work. For the result to be relevant, the setup would however have to include more realistic aspects, such as non-trivial action to state change mapping or more sophisticated agent models (e.g. capable of planification).

6 Conclusion

In this article we studied aspects of the combinatorial structure of human behaviors and of their representation as tasks or objectives. We introduced an algorithm to learn a dictionary of primitive tasks from demonstrations of concurrently mixed behaviors. We demonstrated on an illustrative experiment how the dictionary can be used to represent and generalize new demonstrations. Finally we discussed how dealing with ambiguities in factorial representation of behaviors might involve social interactions, multimodality of the sensory experience or intrinsic saliency mechanisms.

References

1. Konczak, J.: On the notion of motor primitives in humans and robots. In: Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, vol. 123, pp. 47–53. Lund University Cognitive Studies (2005)
2. Kulic, D., Nakamura, Y.: Incremental Learning of Human Behaviors using Hierarchical Hidden Markov Models. In: IEEE International Conference on Intelligent Robots and Systems, pp. 4649–4655. IEEE Computer Society Press (2010)
3. Kruger, V., Herzog, D., Baby, S., Ude, A., Kragic, D.: Learning actions from observations. Robotics and Automation Magazine 17(2), 30–43 (2010)
4. Calinon, S., D'Halluin, F., Sauser, E.L., Caldwell, D.G., Billard, A.G.: An approach based on Hidden Markov Model and Gaussian Mixture Regression. IEEE Robotics and Automation Magazine 17(2), 44–54 (2010)
5. Butterfield, J., Osentoski, S., Jay, G., Jenkins, O.C.: Learning from Demonstration using a Multi-valued Function Regressor for Time-series Data. In: International Conference on Humanoid Robots, vol. (10). IEEE Computer Society Press, Nashville (2010)
6. Paatero, P., Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5(2), 111–126 (1994)
7. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization.. Nature 401(6755), 788–791 (1999)
8. Jenatton, R., Mairal, J., Obozinski, G., Bach, F.: Proximal methods for sparse hierarchical dictionary learning. In: Proceedings of the International Conference on Machine Learning, ICML (2010)
9. Ding, C., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1), 45–55 (2010)

10. Li, Y., Fermuller, C., Aloimonos, Y., Ji, H.: Learning shift-invariant sparse representation of actions. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2630–2637. IEEE, San-Francisco (2010)
11. Hellbach, S., Eggert, J.P., Körner, E., Gross, H.-M.: Basis Decomposition of Motion Trajectories Using Spatio-temporal NMF. In: Alippi, C., Polycarpou, M., Panayiotou, C., Ellinas, G. (eds.) ICANN 2009, Part II. LNCS, vol. 5769, pp. 804–814. Springer, Heidelberg (2009)
12. Mangin, O., Oudeyer, P.-Y.: Learning to recognize parallel combinations of human motion primitives with linguistic descriptions using non-negative matrix factorization. To Appear in International Conference on Intelligent Robots and Systems (IROS 2012), Vilamoura, Algarve (Portugal), IEEE/RSJ (2012)
13. Schaal, S., Ijspeert, A.J., Billard, A.G.: Computational approaches to motor learning by imitation. Phil. Transactions of the Royal Society of London B: Biological Sciences 358, 537–547 (2003)
14. Ratliff, N.D., Bagnell, J.A., Zinkevich, M.A.: Maximum margin planning. In: International conference on Machine learning, ICML 2006, vol. (23), pp. 729–736. ACM Press, New York (2006)
15. Lopes, M., Melo, F., Montesano, L.: Active Learning for Reward Estimation in Inverse Reinforcement Learning. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part II. LNCS, vol. 5782, pp. 31–46. Springer, Heidelberg (2009)
16. Abbeel, P., Coates, A., Ng, A.Y.: Autonomous Helicopter Aerobatics through Apprenticeship Learning. The International Journal of Robotics Research 29(13), 1608–1639 (2010)
17. Jetchev, N., Toussaint, M.: Task Space Retrieval Using Inverse Feedback Control. In: Getoor, L., Scheffer, T. (eds.) International Conference on Machine Learning, ICML 2011, pp. 449–456. ACM Press, New York (2011)
18. Brillinger, D.R.: Learning a Potential Function From a Trajectory. IEEE Signal Processing Letters 14(11), 867–870 (2007)
19. Lefèvre, A., Bach, F.R., Févotte, C.: Online algorithms for Nonnegative Matrix Factorization with the Itakura-Saito divergence. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–9. IEEE Computer Society Press (2011)
20. Massera, G., Tuci, E., Ferrauto, T., Nolfi, S.: The Facilitatory Role of Linguistic Instructions on Developing Manipulation Skills. IEEE Computational Intelligence Magazine 5(3), 33–42 (2010)

Internal Simulations for Behaviour Selection and Recognition

Guido Schillaci[1], Bruno Lara[2], and Verena V. Hafner[1]

[1] Cognitive Robotics Group, Department of Computer Science,
Humboldt-Universität zu Berlin, Germany
{guido.schillaci,hafner}@informatik.hu-berlin.de,
[2] Cognitive Robotics Group, Faculty of Science,
Universidad Autonoma del Estado de Morelos,
Cuernavaca, Mexico
bruno.lara@uaem.mx

Abstract. In this paper, we present internal simulations as a methodology for human behaviour recognition and understanding. The internal simulations consist of pairs of inverse forward models representing sensorimotor actions. The main advantage of this method is that it both serves for action selection and prediction as well as recognition. We present several human-robot interaction experiments where the robot can recognize the behaviour of the human reaching for objects.

Keywords: behaviour recognition, internal simulation, human-robot interaction, internal models.

1 Introduction

Understanding human behaviour is crucial for humans that are inherently social creatures. There has been a strong evolutionary pressure to quickly recognise and understand the actions of others and ideally to be able to map these actions on the own body scheme and experience. The skill of understanding others' behaviour, desires and intentions is closely related to the development of a theory of mind in children around the age of 3.5 years and can be impaired in conditions such as autism [2].

All forms of human behaviour understanding rely on or are based on information that is gained through the observation of others' actions through sensory perception and can be extended with further interaction (sensorimotor and social). In order to build artificial systems that are capable of understanding human behaviour, we aim at understanding the principles of function in humans so they can be transferred to artificial systems.

There is a vast range of potential applications. Human-Robot interaction is an area that would naturally benefit from robots that can understand human behaviour and thus allow for a more intuitive interaction [17]. But also other

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 148–160, 2012.

application areas such as service and surveillance benefit from an automatic detection and recognition of normal and abnormal behaviour. Most systems are based on visual information from cameras that can infer from these the underlying desire, need, mood and intention of a human. Other systems are based on additional sensors, such as acceleration sensors attached to a person, and rely on the particular motion associated with a certain behaviour.

A particular interesting example of behaviour recognition are the experiments with point light walkers by N. Troje and colleagues [20]. They showed that a person can be easily recognized just by observing a small number of light points attached to the joints of the person. Not only the behaviour, but also high-level information such as gender, mood and weight can be extracted.

In previous work [10], we showed that the information can be reduced to the 3-axis acceleration of these points by using an experimental setup where acceleration sensors had been attached to a person. It was possible to recognize different persons and different gaits by analyzing the data from the acceleration sensors. The gaits consisted of walking normally, backwards, upstairs, downstairs and running.

If we want to not only recognize but also to understand human behaviour, we need to be able to map it on our own body scheme and experience. This is closely related to imitation learning where an observed behaviour has to be executed by ourselves. In the next sections, we introduce the concept of internal models which set, we believe, an important ground for performing complex behaviour understanding and imitation.

2 Internal Models

The idea of understanding and studying cognition based on the sensorimotor capabilities of agents is becoming gradually accepted as a research framework [3,15]. In this context, forward and inverse models become central players, as they naturally fuse together sensory and motor information, providing agents with multimodal representations [23,25]. Due to its functioning, these models allot agents with internal simulations, anticipation and predictions, a fundamental basis of cognitive systems.

Forward models were first proposed in the control literature as means to overcome problems such as the delay of feedback on standard control strategies and the presence of noise, a characteristic of natural systems [12]. A forward model is an internal model which incorporates knowledge about sensory changes produced by self-generated actions of an agent. Given a sensory situation S_t and a motor command M_t (intended or actual action) the forward model predicts the next sensory situation S_{t+1}. While forward models (or predictors) present the causal relation between actions and their consequences, inverse models (or controllers) perform the opposite transformation providing a system with the necessary motor command (M_t) to go from a current sensory situation (S_t) to a desired one (S_{t+1}).

In the cognitive sciences, these processes have been found to be capable of modeling several behaviours, ranging from the cancellation of the tickling sensation [4] to the accounting for schizophrenia [8]. The importance of this type of models stems from the relevance of the prediction of the consequences of our actions for seemingly trivial tasks such as planning or avoiding undesired situations [5].

The recent discovery of mirror neurons in the central nervous system supports the general idea of internal simulations. The mirror neuron system (MNS) is thought to be involved in internal simulations of the sensorimotor loop in learning and planning, as it has been found that neurons in this area show activation both when an individual performs a specific action and when the individual observes the same action performed by a demonstrator (for a recent review, see [9]).

It seems that an observer understands a demonstrated behaviour comparing a simulated execution of it with a set of primitives stored in its memory. Here, *Simulation* is seen as the re-enactment of perceptual, motor and introspective states acquired during experience with the world, body and mind [22,3].

Van der Wel et al. [21] view internal models as a mechanism for allowing a faster and more precise interaction with the environment and other agents than would be possible on feedback alone. This is mainly due to the models' prediction capabilities of the perceptual consequences of an action. If people simulate others' actions, then how accurately an observer can predict an observer action should depend on how closely the action maps onto the observer's own motor repertoire.

Much research has been done on computational internal models for action preparation and movement [26], with highly functional models that account, for example, for hand trajectory planning taking into account different contexts [11,24]. These models are also used to set the ground for action recognition and action imitation.

In cognitive robotics, internal models have been used for the execution and recognition of actions [6] and to plan navigation strategies avoiding undesired situations [14,13].

Several architectures have been suggested which are based on the ideas of internal models such as MOSAIC [23], HAMMER [7] or the system presented by Akgun in [1]. The system we report here is based on the ideas reported in [18] and presents in our view a significant difference, namely, the low level control of the agents' movements is based on learning the models through motor babbling. A main advantage is that it requires and works using small displacements of the arm, instead of whole trajectories which would complicate the use for online recognition.

These models have been successfully implemented on a humanoid robot. We are concerned with exploiting and investigating the full potential that inverse-forward models present. First we have conducted experiments on the most basic behaviour, namely action execution, then we use the models to distinguish between two apparently similar actions while reaching a position on space. This behaviour requires two pairs of these inverse-forward models. The next obvious step is now the recognition of actions when not executed by the agent itself. The

arrangement of forward-inverse models presented here for action recognition is the next building block in what we believe to be the scaffolding for cognition.

3 Experiments

In [18], we implemented a mechanism for behaviour selection using internal simulations. A humanoid robot learned a repertoire of behaviours by self-exploration. Each of these sensorimotor schemes was coded as an inverse (controller) - forward (predictor) model pair. The simulated outcomes of each know sensorimotor action have been used for selecting the best strategy from the repertoire to reach a desired goal.

Simulating the outcome of a sensorimotor behaviour consists of two steps. First, an inverse model predicts the motor command necessary to reach a desired sensory situation, according to the sensorimotor behaviour it is coding. Then, an efferent copy of this command is sent to the coupled forward model, for anticipating the sensory situation which would have resulted from the application of that motor command.

In the following subsections, we will show how the same mechanisms can be applied in behaviour and target recognition.

3.1 Behaviour Recognition

Adopting the internal simulation paradigm can solve the task of understanding a human behaviour. During an action demonstration, the actual sensory situation can be compared with the ones predicted by simulating each known sensorimotor scheme stored in the action repertoire. Errors in prediction can be used for classifying the behaviour.

In [18], we reported the use of a computational model for a behaviour recognition experiment. To show the performance of the system, we trained three inverse-forward model pairs. Each of these pairs coded for a different action (see Fig. 7), namely: reach an object; displace the object; withdraw the hand from the object. This set has been chosen because such actions can be described by the variation of the relationships between the position of the hand and the position of the object.

In particular, they have been described with the following characteristics:

- d: distance between hand and object (their 3D positions are estimated using fiducial markers);
- δ: derivative of the distance between hand and object;
- orientation of the object with respect to the hand. This characteristic is coded as two angles, θ and ϕ, which represent the latitude and the longitude of the object position in a frame of reference centered on the hand position.

Each of the three actions is characterised as a different tendency in the variations of such features. For example, the reach action is characterised by a decrease of

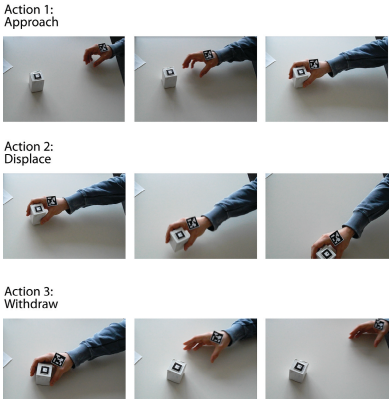


Fig. 1. Sequences of three actions (approach, displace and withdraw) performed by a human subject towards an object

the hand-object distance, thus a negative derivative of the distance. The withdraw action follows the opposite tendency: increase of the hand-object distance, thus positive derivative of the distance. The displace action is characterised by a constant value of the hand-object distance.

In this experiment, the sensory situation S_t is coded as an instance of the previous characteristics: d_t, δ_t, θ_t and ϕ_t. For each time step, the characteristics encoding such a sensory situation are calculated from the positions of the hand and the object. The motor command M_t is coded as the three components of the velocity vector describing the movement of the hand: v^x, v^y and v^z. Table 1 illustrates the input and the output of each internal model in the behaviour recognition experiment.

Table 1. Input and output of the internal models

Inverse Model		
Input	S_{t-1}	$: d_{t-1}, \delta_{t-1}, \theta_{t-1}, \phi_{t-1}$
	S_t	$: d_t, \delta_t, \theta_t, \phi_t$
Output	M_{t-1}	$: v^x_{t-1}, v^y_{t-1}, v^z_{t-1}$

Forward Model		
Input	S_{t-1}	$: d_{t-1}, \delta_{t-1}, \theta_{t-1}, \phi_{t-1}$
	M_{t-1}	$: v^x_{t-1}, v^y_{t-1}, v^z_{t-1}$
Output	S_t	$: d_t, \delta_t, \theta_t, \phi_t$

Supervised learning sessions were performed offline by using recorded videos. The robot observed demonstrations of each action, manually segmented by the user. For each video, data represented by the characteristics specified before, were collected in a knowledge base. Each component of the knowledge base,

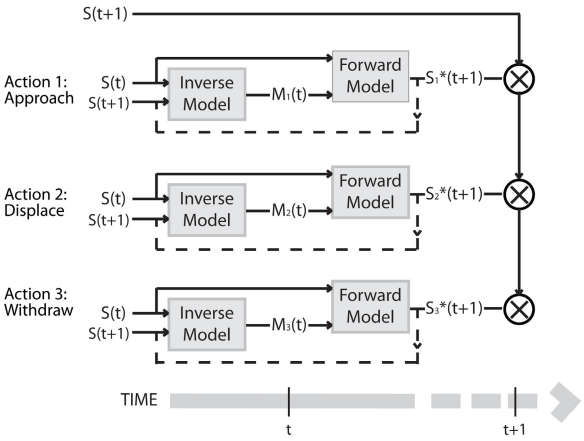


Fig. 2. Inverse (controller) - forward (predictor) model pairs, one pair for each of the three actions

collected at time t, contains the following information: $[S_{t-1}; M_{t-1}; S_t]$, which means that at each time step the previous sensory situation S_{t-1}, the current sensory situation S_t and the motor command M_t that caused S_{t-1} to become S_t have been saved as an element of the knowledge base.

In [18], a k-Nearest Neighbours based algorithm was used as inference tool for the inverse and forward predictions. For inverse model predictions, the motor command M_t (that is, the derivative of the hand displacement) which changes the sensory situation from S_{t-1} to S_t is calculated as follows: Given the hand and object positions at time $t-2$ and $t-1$, the features which compose the sensory situation S_{t-1}, i.e. d_{t-1}, δ_{t-1}, θ_{t-1} and ϕ_{t-1}, are calculated[1]. In a similar way, given the hand and object positions at time $t-1$ and t, the features which compose the sensory situation S_t, i.e. d_t, δ_t, θ_t and ϕ_t, are calculated. A k-NN search in the knowledge base is then performed, where the query is composed by S_{t-1} and S_t. Finally, the M_{t-1} components, i.e. v^x_{t-1}, v^y_{t-1} and v^z_{t-1}, are extracted from the k found vectors and their mean is the output of the inverse model prediction. Figure 3 illustrates the algorithm for performing inverse predictions using k-NN.

Similarly, the forward model predictions are calculated as follows: Given the hand and object positions at time $t-2$ and $t-1$, the features which compose the sensory situation S_{t-1}, i.e. d_{t-1}, δ_{t-1}, θ_{t-1} and ϕ_{t-1}, are calculated. Then, given the hand positions at time $t-2$ and $t-1$, the motor command M_{t-1} is calculated as the derivative of the displacements in each direction, i.e. v^x_{t-1}, v^y_{t-1} and v^z_{t-1}. A k-NN search is performed, but now the query is composed by S_{t-1} and M_{t-1}. The final step consists in extracting from the k found vectors the

[1] Hand and object positions at time $t-2$ are needed for estimating the sensory situation S_{t-1} (for example the variation of the hand-object distance between the current instant and the previous one).

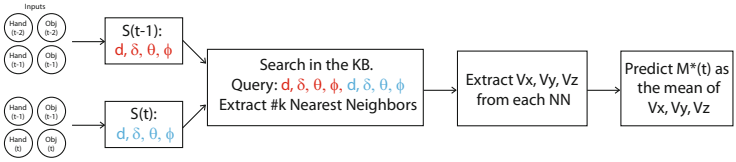


Fig. 3. Illustration of the inverse model prediction with k-NN

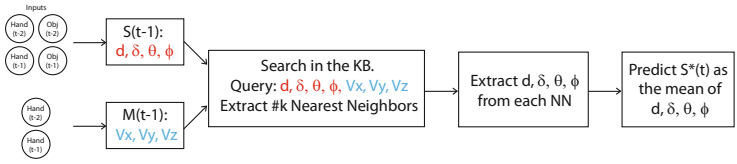


Fig. 4. Illustration of the forward model prediction with k-NN

S_t components, i.e. d_t, δ_t, θ_t and ϕ_t, and returning their mean as the forward model prediction. Figure 4 illustrates the algorithm for performing forward predictions using k-NN.

In the action recognition experiment, the robot is facing towards an action demonstration and is expected to recognise the observed action in real time. Frame by frame, it estimates hand and object positions and it computes both sensory states S_{t-1} and S_t. Internal simulations of the sensorimotor loop are performed for each action, that is for each controller-predictor pair. First, S_{t-1} and S_t are fed into the inverse model which predicts the motor command M^*_{t-1}; then, S_{t-1} and M^*_{t-1} are sent to the corresponding forward model to generate the simulated outcome S^*_t. Each of these predictions is then compared with the actual sensory situation S_t. The action corresponding to the pair with the least error is chosen as the most probably observed one[2].

Preliminary results of the behaviour recognition experiment using the k-NN based inference algorithm have been partially presented in [19].

In this work we present the performance of the behaviour recognition system with a different inference tool. Here, each inverse and forward model has been coded as a multi-layer perceptron which has been trained with a backpropagation algorithm using the data collected during the supervised learning sessions. The number of inputs, outputs and hidden nodes varied according to the trained model: for the forward model, the input neurons get the values for S_{t-1}, i.e. d_{t-1}, δ_{t-1}, θ_{t-1} and ϕ_{t-1}, and for M_{t-1}, i.e. v^x_{t-1}, v^y_{t-1} and v^z_{t-1}, and the output neurons code S_t, i.e. d_t, δ_t, θ_t and ϕ_t; for the inverse model, the input neurons get the values for S_{t-1}, i.e. d_{t-1}, δ_{t-1}, θ_{t-1} and ϕ_{t-1}, and for S_t, i.e. d_t, δ_t, θ_t and ϕ_t, and the output neurons code M_{t-1}, i.e. v^x_{t-1}, v^y_{t-1} and v^z_{t-1}.

[2] Mahalanobis distance has been used for the comparisons.

3.2 Target Recognition

In the behaviour recognition experiment, we performed internal simulations in order to understand which inverse-forward model pair was more closely coding for the observed demonstration. In this section, we show how this system can be used to recognize not only the action performed on an object, but also the target object of the action, in case there are several objects in the scene.

As described before, the sensory states S_{t-1} and S_t of the internal models correspond to certain relationships between the position of the hand of the demonstrator and the one of the target object. During the demonstration, S_{t-1} and S_t are extracted from such positions and sent into the inverse and forward model to generate motor and state predictions. Each state prediction S_t^* (that is, the outcome of each controller-predictor pair) is then compared to the actual one, S_t. The observed behaviour is then classified as the one corresponding to the pair which results in the lowest prediction error.

When there is more than one object in the scene, simulations can be performed for each object. The same internal models can be fed with the states computed using the relationship between the position of the hand and each one of the objects, for example S_{t-1}^1 (i.e. d_{t-1}^1, δ_{t-1}^1, θ_{t-1}^1 and ϕ_{t-1}^1) and S_t^1 (i.e. d_t^1, δ_t^1, θ_t^1 and ϕ_t^1) as the states computed with the position of object 1, S_{t-1}^2 and S_t^2 as the states computed with the position of object 2, etc. Thus, we can feed each inverse-forward model pair with each of these couples of states and compute the prediction errors with their corresponding desired states, in such a way that we can infer the target object of the ongoing action as the one which corresponds to the best inverse-forward model pair fed with the states computed with its position.

Assume that we have two objects, 1 and 2, and two inverse-forward model pairs (the first coding for the *approach* action and the second coding for the *displace* action). The system computes the states S_{t-1}^1 and S_t^1 (using the position of the hand and the one of object 1) and the states S_{t-1}^2 and S_t^2 (using the position of the hand and the one of object 2). S_{t-1}^1 and S_t^1 are sent to the pair *approach*, a prediction S_t^{1*} is calculated and compared with the state S_t^1, resulting in the prediction error $ERR_{approach}^1$. In the same way, S_{t-1}^2 and S_t^2 are sent to the pair *approach*, resulting in the prediction error $ERR_{approach}^2$. The same process is done with the pair *displace*, resulting in two more prediction errors, so that in total we have: $ERR_{approach}^1$, $ERR_{approach}^2$, $ERR_{displace}^1$ and $ERR_{displace}^2$. The smallest error corresponds to the best pair which is fed with the data of the most probable target of the action.

In the next section, we will show quantitative results of the performance of the behaviour recognition system together with the target recognition one, using two different inference tools (k-NN and Multi-Layer Perceptron).

4 Results

In this section, we present the results of the behaviour and target recognition system using internal simulations.

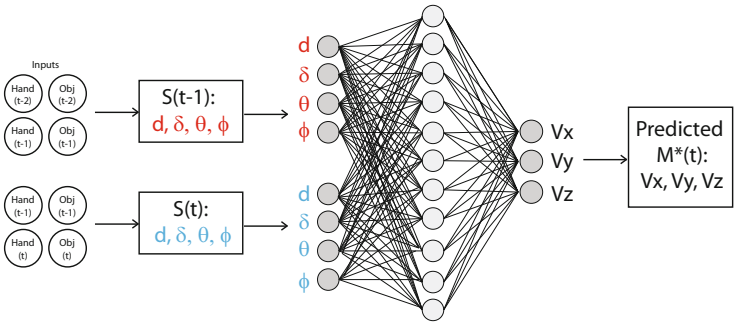


Fig. 5. Illustration of the inverse model prediction with MLP

We trained two inverse-forward models pairs with data collected from the observation of two actions directed to an object: *approach* and *displace*. For the first one, we recorded 1004 samples taken from 83 demonstrations; for the second one, we collected 3245 samples from 108 demonstrations. Each sample contains $[S_{t-1}; M_{t-1}; St]$, where the state S is a 4-dimensional vector containing the same features described in section 3.1: d, δ, θ and ϕ. As before, the motor command M is a 3-dimensional vector representing the x,y,z velocities of the hand.

Two tools have been used for training the models and performing predictions: k-Nearest Neighbours and Multi-Layer Perceptrons. The former has been described in [16] and [18]. Regarding the multi-layer perceptrons, we trained a neural network for each internal model in each pair. The training parameters were the same for all of them[3]. The forward models have been coded as MLPs with 7 input neurons, 12 neurons in the hidden layer, and 4 output neurons. The inverse models have been represented as MLPs with 8 input neurons, 16 neurons in one hidden layer and 3 output neurons[4]. In training the internal models for the *approach* behaviour, the epsilon threshold term criteria has been reached after 437 iterations for the forward model and after 2136 iterations for the inverse one. In the *displace* case, the epsilon threshold term criteria has been reached after 333 iterations for the forward model and after 1891 iterations for the inverse one. Figure 5 and 6 illustrate the algorithms for inverse and forward predictions using MLPs.

The following tables show the confusion matrices using four inference tools: MLP (Multi-Layer Perceptron), 5-NN (k-Nearest Neighbours with $k = 5$), 11-NN and 55-NN. A confusion matrix indicates how much (in percentiles of the

[3] Term criteria: MaxIteration=500000; Epsilon= 0.000001; Activation function = Symmetrical Sigmoid; Training algorithm = BackPropagation; dw-scale (the coefficient to multiply the computed weight gradient by) = 0.05; moment-scale (the coefficient to multiply the difference between weights on the 2 previous iterations. This parameter provides some inertia to smooth the random fluctuations of the weights) = 0.05

[4] 4 input neurons to code for S_{t-1} plus 4 input neurons for S_t and 3 output neurons for M_{t-1}.

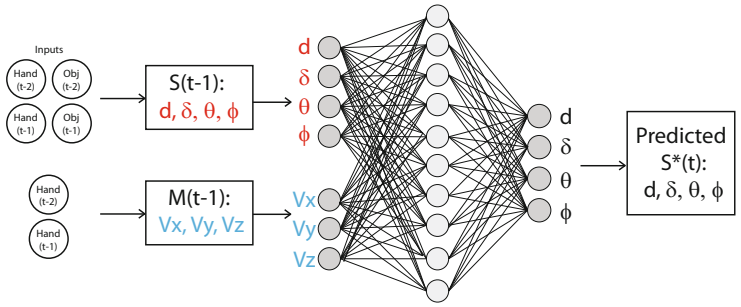


Fig. 6. Illustration of the forward model prediction with MLP

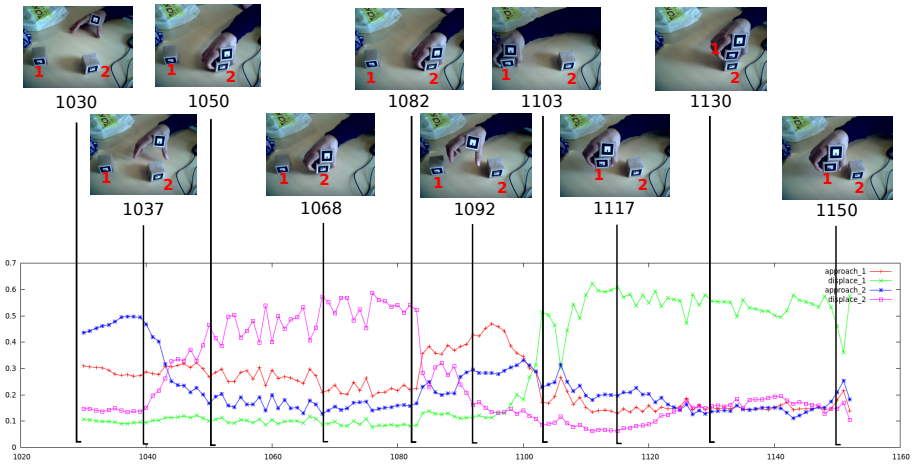


Fig. 7. Demonstration of the behaviour and target recognition using MLP. The bottom graph shows the probabilities of each action to each object.

Table 2. Confusion Matrix. MLP.

		Actual Outcome			
		approach-obj1	displace-obj1	approach-obj2	displace-obj2
Prediction Outcome	approach-obj1	100.00%	0.00%	24.45%	7.25%
	displace-obj1	0.00%	95.52%	0.00%	0.00%
	approach-obj2	0.00%	4.48%	73.33%	0.00%
	displace-obj2	0.00%	0.00%	2.22%	92.75%

actual outcome) every demonstrated action has been recognised as approach or displace with target object 1 or 2. The correct classification rates were: 89.45% for the MLP, 61.81% for the 5-NN, 63.82% for the 11-NN and 64.82% for the 55-NN, claiming MLP as the best performing tool.

Table 3. Confusion Matrix. k-NN ($k = 5$).

		Actual Outcome			
		approach-obj1	displace-obj1	approach-obj2	displace-obj2
Prediction Outcome	approach-obj1	70.59%	0.00%	57.78%	39.13%
	displace-obj1	0.00%	97.01%	4.44%	0.00%
	approach-obj2	29.41%	2.99%	8.89%	0.00%
	displace-obj2	0.00%	0.00%	28.89%	60.87%

Table 4. Confusion Matrix. k-NN ($k = 11$).

		Actual Outcome			
		approach-obj1	displace-obj1	approach-obj2	displace-obj2
Prediction Outcome	approach-obj1	70.59%	0.00%	55.56%	33.33%
	displace-obj1	0.00%	95.52%	2.22%	0.00%
	approach-obj2	29.41%	4.48%	11.11%	0.00%
	displace-obj2	0.00%	0.00%	31.11%	66.67%

Table 5. Confusion Matrix. k-NN ($k = 55$).

		Actual Outcome			
		approach-obj1	displace-obj1	approach-obj2	displace-obj2
Prediction Outcome	approach-obj1	70.59%	0.00%	51.11%	31.88%
	displace-obj1	0.00%	97.01%	2.22%	0.00%
	approach-obj2	29.41%	2.99%	11.11%	0.00%
	displace-obj2	0.00%	0.00%	35.56%	68.12%

5 Conclusions

Grounded theories of cognition support the idea that knowledge relies on the connections between modal representations of action, perception and introspection. Simulation, intended as the process of re-enactement of previously experienced motor, perceptual or introspective situations, has become a central capability and requirement for cognition.

We showed how internal simulations of the sensorimotor loop can be used in understanding a human motor behaviour and highlighted the internal models representation as a good candidate for encoding sensorimotor schemes and for performing simulations.

We believe that the paradigm of internal simulations as a mechanisms for understanding (if we consider *understanding* as that process of linking abstract symbols with real or simulated modal representations) can be applied in different contexts and levels of abstraction.

As mentioned in Section 2 we are concerned with the study of inverse-forward models and their potential as the grounding of cognition in agents. We would like to focus our efforts and future work on using these models for addressing more low level behaviours such as the recognition of self when executing an action versus someone else's execution, even when these observations are performed off-line. Although apparently much has been done and reported with many different and interesting architectures we believe that many questions remain open with regards to the full capabilities these models allot agents on their dealing with the environment.

Acknowledgments. This work has been partially financed by the EU funded Initial Training Network (ITN) in the Marie-Curie People Programme (FP7): INTRO (INTeractive RObotics research network), grant agreement no.: 238486. The work of Prof. Lara is funded by the Alexander von Humboldt Foundation through a Georg Forster Fellowship.

References

1. Akgün, B., Tunaöglu, D., Sahin, E.: Action recognition through an action generation mechanism. In: International Conference on Epigenetic Robotics (EPIROB) (2010)
2. Baron-Cohen, S.: Mindblindness: An Essay on Autism and Theory of Mind. MIT Press (2001)
3. Barsalou, L.W.: Grounded cognition. Annual Reviews Psychology 59, 617–645 (2008)
4. Blakemore, S.J., Wolpert, D., Frith, C.: Why can't you tickle yourself? Neuroreport 11, 11–16 (2000)
5. Blakemore, S.J., Goodbody, S.J., Wolpert, D.M.: Predicting the consequences of our own actions: The role of sensorimotor context estimation. The Journal of Neuroscience 18(18), 7511–7518 (1998)
6. Dearden, A.: Developmental learning of internal models for robotics. Ph.D. thesis, Imperial College London (2008)
7. Demiris, Y., Simmons, G.: Perceiving the unusual: Temporal properties of hierarchical motor representations for action perception. Neural Networks pp. 272–284 (2006)
8. Frith, C.D.: The Cognitive Neuropsychology of Schizophrenia. Erlbaum Associates (1992)
9. Gallese, V.: Before and below theory of mind: embodied simulation and the neural correlates of social cognition. Phil. Trans. of the Royal Society B 362(1480), 659–669 (2007)
10. Hafner, V.V., Bachmann, F.: Human-humanoid walking gait recognition. In: Proceedings of Humanoids 2008, 8th IEEE-RAS International Conference on Humanoid Robots, pp. 598–602 (2008)
11. Haruno, M., Wolpert, D.M., Kawato, M.: Mosaic model for sensorimotor learning and control. Neural Computation 13, 2201–2220 (2001)
12. Jordan, M.I., Rumelhart, D.E.: Forward models: Supervised learning with a distal teacher. Cognitive Science 16, 307–354 (1992), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.657

13. Lara, B., Rendon, J.M., Capistran, M.: Prediction of multi-modal sensory situations, a forward model approach. In: Proceedings of the 4th IEEE Latin America Robotics Symposium, vol. 1 (2007)
14. Möller, R., Schenck, W.: Bootstrapping cognition from behavior–a computerized thought experiment. Cognitive Science 32(3), 504–542 (2008), http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ799248
15. Prinz, W.: Perception and action planning. European Journal of Cognitive Psychology 9(2), 129–154 (1997)
16. Schillaci, G., Hafner, V.V.: Random movement strategies in self-exploration for a humanoid robot. In: Proc. of the Intern. Conf. on Human-Robot Interaction 2011, pp. 245–246 (2011)
17. Schillaci, G., Hafner, V.V.: Prerequisites for intuitive interaction - on the example of humanoid motor babbling. In: HRI 2011 Workshop on Expectations in intuitive human-robot interaction, Laussane, Switzerland (March 2011)
18. Schillaci, G., Hafner, V.V., Lara, B.: Coupled inverse-forward models for action execution leading to tool-use in a humanoid robot. In: Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction, Boston (2012)
19. Schillaci, G., Lara, B., Hafner, V.V.: Internal simulation of the sensorimotor loop in action execution and recognition. In: Proceedings of the 5th International Conference on Cognitive Systems (CogSys 2012), Vienna, Austria (2012)
20. Troje, N.F.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of Vision 2(5), 371–387 (2002)
21. van der Wel, R., Sebanz, N., Knoblich, G.: Action perception from a common coding perspective. In: Johnson, K., Shiffrar, M. (eds.) People Watching: Social, Perceptual, and Neurophysiological Studies of Body Perception (in press)
22. Wilson, M., Knoblich, G.: The case for motor for motor involvement in perceiving conspecifics. Psychological Bulletin 131, 460–473 (2005)
23. Wolpert, D.M., Kawato, M.: Multiple paired forward and inverse models for motor control. Neural Netw. 11(7-8), 1317–1329 (1998)
24. Wolpert, D.M., Doya, K., Kawato, M.: A unifying computational framework for motor control and social interaction. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 358(1431), 593–602 (2003), http://dx.doi.org/10.1098/rstb.2002.1238
25. Wolpert, D.M., Flanagan, J.R.: Motor prediction. Current Biology 11(18), R729–R732 (2001),
26. Wolpert, D.M., Ghahramani, Z.: Computational principles of movement neuroscience. Nature Neuroscience 3(suppl.), 1212–1217 (2000),

Automatic Imitation Assessment in Interaction

Stéphane Michelet*, Koby Karp*, Emilie Delaherche,
Catherine Achard, and Mohamed Chetouani

Institute of Intelligent Systems and Robotics,
University Pierre and Marie Curie, 75005 Paris, France
{michelet,karp,delaherche}@isir.upmc.fr,
{catherine.achard,mohamed.chetouani}@upmc.fr
http://www.isir.upmc.fr/

Abstract. Detecting social events such as imitation is identified as key step for the development of socially aware robots. In this paper, we present an unsupervised approach to measure immediate synchronous and asynchronous imitations between two partners. The proposed model is based on two steps: detection of interest points in images and evaluation of similarity between actions. Firstly, spatio-temporal points are detected for an accurate selection of the important information contained in videos. Then bag-of-words models are constructed, describing the visual content of videos. Finally similarity between bag-of-words models is measured with dynamic-time-warping, giving an accurate measure of imitation between partners. Experimental results obtained show that the model is able to discriminate between imitation and non-imitation phases of interactions.

Keywords: Imitation, DTW, unsupervised learning.

1 Introduction

Face-to-face interactions are considered as highly dynamic processes [1,2] based on multimodal exchanges such as turn-taking, backchannels (e.g., head nod, filled pauses...). Sensing, characterizing and modeling interactions are challenging. Various natures of human communication dynamics have to be taken into account: individual (e.g., gesture completion), interpersonal (e.g., mimicking)... In recent years, there has been a growing interest for human communication dynamics in several domains such as Social Signal Processing and Social Robotics. In [3,4], backchannels are investigated firstly by modeling human-human communication and then the model is employed to generate multimodal feedbacks by an agent (Embodied Communicative Agents/ Robots) during dialogs. Thanks to these dynamical models, agents are able to provide relevant communicative responses and consequently to sustain interactions. Continuously monitoring social exchanges between partners is a fundamental step of social robotics [5].

* These authors contributed equally.

A.A. Salah et al. (Eds.): HBU 2012, LNCS 7559, pp. 161–173, 2012.

Interpersonal dynamics, usually termed interpersonal synchrony [1] is a very complex phenomenon including various concepts such as imitation, mimicking, turn-taking... In this paper, we focus on immediate imitation characterization, which include synchronous and asynchronous reproductions of a demonstrated action within few seconds. Here, imitation is considered as a communicative act allowing to sustain interaction [1]. Being able to automatically assess imitation between social partners (including agents) is required for developing socially intelligent robots [6,1]. Given this framework, the proposed method is seen to be different to traditional approaches for learning from demonstration in human-robot interaction [7], where imitation metrics usually assess kinematic, dynamic and timing dimensions.

The paper is organized as follows: Section 2 reports recent works on interpersonal synchrony characterization formulated as an action recognition problem. Section 3 briefly describes the approach proposed for imitation assessment. Sections 4 and 5 describe the different steps of our model based on 1) characterization of actions (spatio-temporal interest points, bag-of-words) and 2) similarity metrics (correlation, dynamic time warping). Section 6 presents results on a gesture imitation task. Finally, a conclusion provides a summary of the model discussed throughout the paper and proposes future works.

2 Related Work

Currently, few models have been proposed to capture mimicry in dyadic interactions. Mimicry is usually considered in the larger framework of assessing interactional synchrony, the coordination of movement between individuals in both timing and form during interpersonal communication [8]. Actual state-of-the-art methods to assess synchrony rely on two steps: feature extraction and a measure of similarity.

The first step in computing synchrony is to extract the relevant features of the dyad's motion. We can distinguish between studies focusing on the movement of a single body part and those capturing the overall movement of the dyad. Numerous studies focus on head motion, which can convey emotion, acknowledgement or active participation in an interaction. Head motion is captured using either a motion-tracking device [9] or a video-based tracking algorithm [10,11]. Many studies capture the global movements of the participants with Motion Energy Images [12,13] or derivatives [14,15].

Then, a measure of similarity is applied between the two time series. Correlation is certainly the most commonly used method to evaluate interactional synchrony. After extracting the movement time series of the partners, a time-lagged cross-correlation is applied between the two time series using short windows of interaction. Several studies also use a peak picking algorithm to estimate the time-lag between the partners [9,16,12]. Recurrence analysis is an alternative to correlation [11]. It was inspired by the theory of coupled dynamical systems, providing graphical representations of the dynamics of coupled systems. Recurrence analysis assesses the points in time that two systems show similar patterns of

change or movement, called "recurrence points". These models are often poorly selective for mimicry detection. Indeed, the features (e.g. motion energy) describe rather the amount of movement than the form of the gestures performed. Capturing mimicry entails to have a finer description of the gestures. That can be reached using action recognition techniques.

In the last few years, many researches have emerged in this domain as described in numerous reviews [17], [18], [19]. The first approaches consist to characterize the sequences globally. Davis and Bobick [20] introduced the Motion History Images (MHI) and the Motion Energy Images (MEI) that summarize in a single image all the motions performed during the sequence. Then, simple moments on these images characterize the sequence. In order to preserve movement kinetic, Mokhber et al. [21] proposed to directly characterize the spatio-temporal volume by geometric moments. Efros et al. [22] characterized individually each image thanks to an optical flow based feature, and then compared the sequences with a measure similar to correlation. Laptev et al. [23] explored the combination of local space-time features histograms and SVM. First, spatio-temporal interest points are detected by extending the Harris detector to the space-time domain. These points are then characterized using several motion representations in terms of spatio-temporal jets, position dependent histograms, position independent histograms, and principal component analysis computed for either spatio-temporal gradients or optic flow. Dollár et al. [24] introduced a new spatio-temporal interest points detector explicitly designed to detect more points and to be more robust. They are then described with spatio-temporal cuboids.

Less works have been made on unsupervised action recognition. Niebles et al. [25] represent a video as a collection of spatial-temporal words by extracting space-time interest points. The algorithm automatically learns the probability distributions of the spatial-temporal words and the intermediate topics corresponding to human action categories. This is achieved by using latent topic models such as the probabilistic Latent Semantic Analysis (pLSA). Rao et al. [26] describe a representation of human action that captures changes in speed and direction of the trajectory using spatio-temporal curvature of 2-D trajectory. Starting without a model, they use this representation for recognition and incremental learning of human actions. Zelnik-Manor and Irani [27] design a simple statistical distance measure between video sequences (possibly of different lengths) based on their behavioral content. It is used to isolate and cluster events within long continuous video sequences, without prior knowledge of the types of events, their models, or their temporal extent.

3 Overview of Our Approach

In this paper, we propose an innovative approach to measure imitation between two partners, through the use of unsupervised action recognition. Indeed, instead of characterizing a video with global measures by only quantifying movement, we are considering imitation as the similar execution of actions, in which the semantic of actions is not needed. In order to create an imitation rate, as shown

in Fig. 1, the first step is to detect regions of interest, called Spatio-Temporal Interest Points (STIPs) which will be described in section 4.1. They are then described using local histograms, leading to bag-of-words models. Section 5 will give details on the similarity measure which is applied on these models to compare two videos. Finally, as shown in section 6.2 an imitation rate is fitted from the similarity obtained.

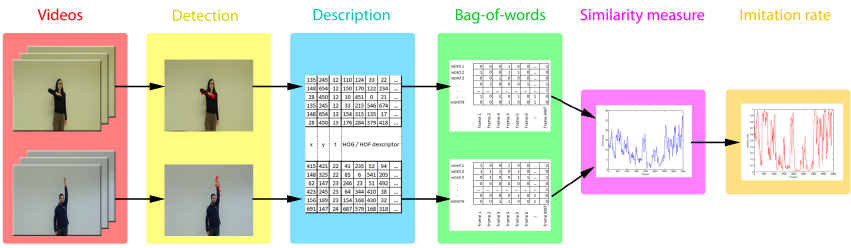


Fig. 1. Video analysis process

4 Modeling Videos with Bag-of-Words

As it is usually not feasible and impractical to measure correlations between all regions of two videos due to computation limitations, we are using methods that will detect significant areas in the video that are rich in information, also known as Spatio-Temporal Interest Points (STIPs).

4.1 Detection

Detection of the STIPs is based on Dollár's work in [24], where detection of low-level features is performed using 1D Gabor filters. Even though Gabor filters were designed to return high response for periodic motions such as a bird flapping its wings, it has proven to evoke strong response for non-periodic elements such as motion of spatio-temporal corners. It has been preferred to other detectors because of its robustness and for its quantity of correctly detected points (around 12 per frame), allowing a good characterization of the video.

4.2 Description

After obtaining local maxima points from Dollár's response function, each point's spatio-temporal neighborhood (patch) is characterized. However, Dollár's descriptor is based on cuboids, which are expensive both in terms of computation and memory. Indeed, the descriptor uses 19-by-19-by-11 cuboids (9 pixels on each sides and 5 frames before and 5 after), which gives a descriptor with dimension 3971 for each point.

Each point has thus been described by HOG/HOF (Histogram of Oriented Gradients/Histogram of Optical Flow) descriptor as introduced by Laptev in [28]. While HOG has strong similarity to the well known SIFT descriptor, HOF is based on occurrences of orientations of optical flow. Each patch is divided into a grid of cells and for each cell 4-bin HOG and 5-bin HOF histograms are then computed and concatenated into a single feature vector. HOG/HOF descriptors return vectors with 162 elements (72 for HOG and 90 for HOF). HOG/HOF dimensionality is more compact and therefore more convenient than Dollar's cuboid for this application.

4.3 Construction of a Vocabulary and Bag-of-Words Model

Similarly to natural language processing, the next step clusters the collection of points in a vocabulary describing the videos. In natural language processing, a cluster would be called a lexical field, thus putting together words like "drive, driving, driver". STIPs are clustered using a k-means technique, forming a k video-words dictionary. The dictionary is formed using training videos.

Then, bag-of-words models are created for each of the two videos. In order to do so, each STIP detected is assigned to the nearest video-word with Euclidean distance. Following representation of each observation as a word in a vocabulary, the entire video is represented as occurrences of words using the bag-of-words model. Every frame is characterized by a histogram of words. This results in a sparse matrix $BOW(w, f)$ of high dimensions : number-of-words by number-of-frames. The bag-of-words models describe the temporal structure of the video, but looses the spatial one. Indeed, the spatial coordinates of a word does not appear in this model, thus the spatial position of a video-word in a frame does not have any impact.

5 Similarity Measures

After describing the videos by bag-of-words models, a similarity measure can then be applied to compare them. The first approach coming to mind is to use correlation. However, due to a delay between the imitation initiator and the imitator and to the very sparse nature of bag-of-words, direct correlation is a poor measure. The first idea we present here is to take enlarged analysis windows on which correlation is computed. But since taking the dynamic of the imitation into account is important in interaction, we applied a modified version of Dynamic Time Warping in which similarity is measured.

5.1 Correlation

To allow small variations in time, we represent each instance as a vector and measure sum of words inside a corresponding window as presented in equation 1, where win is the size of the window, w is a word, and BOW_A is the bag-of-words of video A, $a_f(w)$ is the summed vector that defines video A in each

instance indexed by the number f of the frame it starts from. Given the fact that a gesture that occurs after more than three seconds can not be considered as imitation, a sliding window of 75 frames has been chosen like shown in Fig. 5 (since the video frame rate is 25 frames per second).

$$a_f(w) = \sum_{i=f}^{f+win} BOW_A(w,i) \tag{1}$$

After having the corresponding summed vector for each video, the two time series a_f and b_f are compared using normalized correlation coefficient, as recalled in equation 2.

$$normcorr(a_f, b_f) = \frac{a_f^T(w).b_f(w)}{||a_f(w)||_2 * ||b_f(w)||_2} \tag{2}$$

5.2 Dynamic Time Warping

In natural interaction the time-lag of imitation between partners varies all the time and partners continuously change roles. Thus, a straight similarity measurement like correlation is not able to take into account the variations in the time-delay between partners. Thus a dynamic comparison of the imitation between the partners is needed. In this matter, Dynamic Time Warping is a reference to compare two non aligned time series. However, as we are going to present here, we are not using DTW to measure a distance, but to measure similarity.

Whereas the original method from Levenshtein [29] measures a distance between two series, the one developed by Needleman [30] permits to measure similarity and has been widely used with DNA-strand for genome comparison, or sequence alignment. In this last method, the computation of the cumulated similarity matrix follows equation 3, where *normcorr* refers to the normalized correlation defined in equation 2. Detailed explanations on Dynamic Time Warping can be found in Chapter 4 of [31].

$$D(i,j) = \max \begin{cases} D(i-1,j) - (1 - normcorr[x(i-1), x(i)]) \\ D(i,j-1) - (1 - normcorr[y(j-1), y(j)]) \\ D(i-1,j-1) + normcorr[x(i), y(j)] \end{cases} \tag{3}$$

Once the cumulated similarity matrix is computed, the shorter path is searched for. Usually, as one wants to warp two time series together, the matrix is computed such by warping the whole sequence A with the whole sequence B, as shown in Fig. 2.

But, as illustrated in Fig. 3, the Overlap Detection variant permits to ignore the beginning of one time series and the end of the other one. This is important for imitation measurement as it permits to have gestures from different lengths, surrounded by gestures which does not fulfill imitation. The only modification to the original algorithm is that the first line and the first column of the cumulated similarity matrix are set to zero, and the maximum score is searched in the whole last column and last line.

The second interesting variation, called Local Alignment, permits to ignore a part of the start and the end for each time series. It is called the Smith-Waterman algorithm (refer [32]). This is illustrated in Fig. 4. It permits to compare short gestures in two parts of videos. This algorithm is now modified by adding a fourth option to equation 3, so that D(i,j) is the maximum of either those three options or 0. Then if all the numbers are negative, it "resets" the path, permitting to have a new starting point, and leading to results as shown in Fig. 4. Moreover, the maximum score is now researched in the whole matrix.

During this whole paper, the Local Alignment method has been used.

Fig. 2. Classic DTW

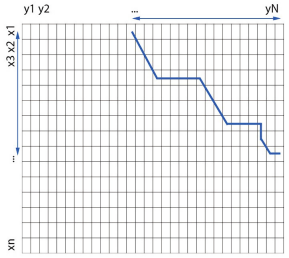


Fig. 3. Overlap DTW

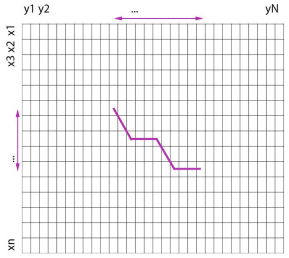


Fig. 4. Local DTW

In order to allow some time-lag between partners, the two analysis windows can have different lengths. From video A, BOW_A is made with a 75-frames window and is compared with BOW_B made by 125-frames windows (75 plus two times 25 frames), thus allowing a time-lag of one second. The two versions of DTW (75-75 and 75-125) will be compared in the Results section, and are shown in Fig. 5.

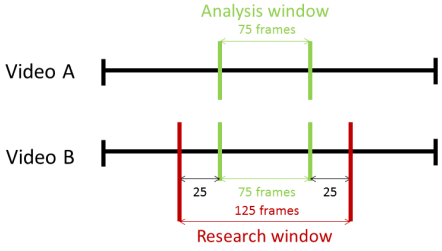


Fig. 5. Windows analysis and research

6 Experiments

6.1 Data

Current databases often enhance the gesture recognition, or give multiple videos of the same action, done in different contexts. However, few data with

synchronized videos are available for interpersonal studies, with annotation. A database of synchronized gestures for two partners has been presented in [33]. Table 1 gives the characteristics of this database, and figures 6 and 7 are illustrations of it.

Table 1. Stimuli and conditions. We denote for each sequence its length l in seconds and the number of gestures n in the sequence l[n].

Frequency (in BPM)	Synchrony and No Imitation (S _NBM)	Synchrony and Imitation (S_BM)
20	137[44]	62[19]
25	166[67]	71[28]
30	153[71]	59[27]

Fig. 6. Imitation dual video

Fig. 7. Non imitation dual video

6.2 Results

Validation of the Protocol: To be sure that the methods actually permits to separate imitation from non-imitation, a first test has been performed with correlation between 75-frames windows. Correlation score is computed at each time for both imitation and non-imitation videos. The results are shown on Fig. 8 for the two classes, non-imitation videos being longer than imitation videos. Distributions of the two series are shown in Fig. 9, and a t-test permits to verify that as seen on this figure, the two statistical series are separable (h=1 with $p < 0.05$). The method is thus suitable for unsupervised imitation measurement.

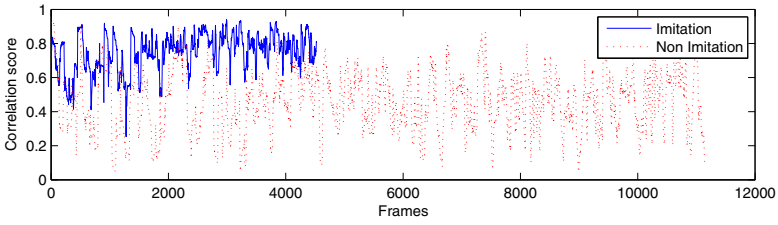


Fig. 8. Correlation score for two imitation videos (solid line) and two non-imitation videos (dotted line)

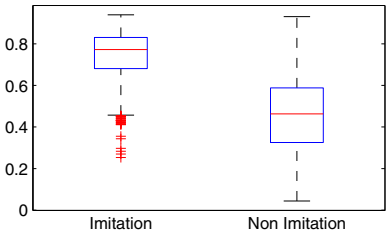


Fig. 9. Distribution of the correlation score

Comparison of Methods: The protocol has been applied to three methods : correlation, DTW 75-75 and DTW 75-125. In order to compare and evaluate the efficiency, Receiver Operating Characteristic (ROC curves) are used. They are estimated using all correlation measures obtained for each time and each video. Results for the three methods are shown in Fig. 10, where the True Positive Rate (TPR) is plotted as a function of the False Positive Rate (FPR). The Area Under Curve (AUC) is often referred as an efficiency measurement of classifiers. However, as it can be seen in the Fig. 10, even if correlation and DTW 75-125 have the same AUC, the curves show better results for DTW 75-125 in the part near the optimal point. This is confirmed by the d' measure, computed by $d' = \max_i [TPR(i) - FPR(i)]$. The d' measure gives the optimal working point, which is significantly higher for DTW 75-125 (Fig. 12). Moreover, one could note that the results for DTW are not highly superior to correlation. This is explained by the structure of the dataset, which has been created in almost perfect synchrony, and thus where dynamic variations are absent, leading to comparable results for correlation and DTW.

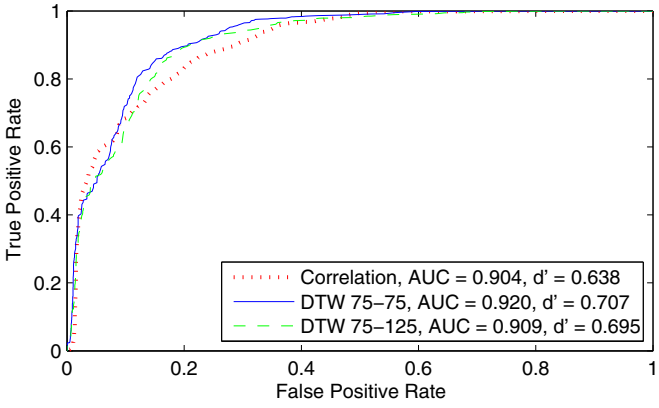


Fig. 10. ROC curves for the three methods

Robustness: As this protocol is aimed to be used in natural interactions, it has to be robust to shifting. In order to evaluate its robustness, tests have been done by shifting one of the sequences temporally (between -1s and 1s, equivalent to +/- 25 frames). A delay of more than one second has not been envisaged as it cannot be seen as a real imitation. Comparisons were made using ROC, but to summarize results, only *AUC* and d' measures are presented in figures 11 and 12. Even if DTW 75-75 and correlation seem to have similar results on AUC curves for negative delay, DTW 75-75 outperforms correlation with d' measures, which better represents the real use of the system (near the operating point). The third method DTW 75-125 is robust to delay between sequences, which has very little influence on the results. Indeed, as shown in Figs. 11 and 12, the results are very stable for shiftings between -25 frames and +25 frames.

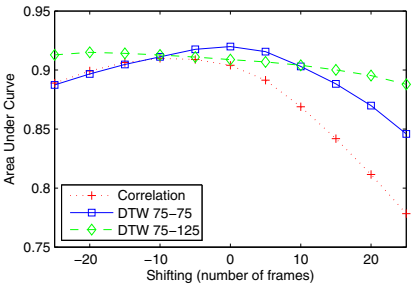


Fig. 11. AUC variations with shifting

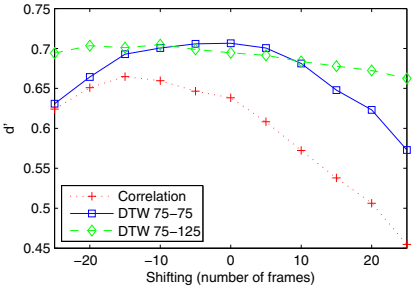


Fig. 12. d' variations with shifting

7 Conclusion

We have proposed in this paper an efficient process to measure imitation rate during an interaction between partners using unsupervised action recognition. However, the database presented here is strongly synchronized, thus giving more advantage to correlation than to DTW. Moreover, no dynamic appears in the data, such as turn-taking, reducing the interest in terms of interpretation of DTW results. The next step will be to test this method on more natural gestures and interactions in which dynamics will play a large role.

However correlation and DTW each have their interest. Correlation, by its fast computation, permits to have good results as long as data is not too much shifted. On the other hand, even if DTW is a bit more computationally expensive, it permits to take into account the dynamics of interaction, and it will be used in further developments.

Moreover, HOG/HOF descriptor does not take into account the spatial localization of the video-words. Adding information on the relative position of the detected points has been envisaged for the future in order to increase accuracy.

Some further tuning of DTW can be done to improve results, which have not been covered here. The Local Alignment method gave slightly better results

than the Overlap Detection, but the influence of the two methods has not been studied here. However, in real interaction videos, some first tests have permitted to see differences between the two algorithms, which will be studied in further developments.

References

1. Delaherche, E., Chetouani, M., Mahdhaoui, M., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: A survey of evaluation methods across disciplines. IEEE Transactions on Affective Computing (to appear, 2012)
2. Morency, L.-P.: Modeling human communication dynamics (social sciences). IEEE Signal Processing Magazine 27(5), 112–116 (2010)
3. Morency, L.-P., de Kok, I., Gratch, J.: Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 176–190. Springer, Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-85483-8_18
4. Al Moubayed, S., Baklouti, M., Chetouani, M., Dutoit, T., Mahdhaoui, A., Martin, J.-C., Ondas, S., Pelachaud, C., Urbain, J., Yilmaz, M.: Generating robot/agent backchannels during a storytelling experiment. In: IEEE International Conference on Robotics and Automation, ICRA 2009, pp. 3749–3754 (2009)
5. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artif. Intell. 166, 140–164 (2005)
6. Rolf, M., Hanheide, M., Rohlfing, K.: Attention via synchrony: Making use of multimodal cues in social learning. IEEE Trans. Auton. Mental Develop. 1(1), 55–67 (2009)
7. Calinon, S., D'halluin, F., Sauser, E., Caldwell, D., Billard, A.: Learning and reproduction of gestures by imitation: An approach based on Hidden Markov Model and Gaussian Mixture Regression. IEEE Robotics and Automation Magazine 17(2), 44–54 (2010)
8. Bernieri, F.J., Reznick, J.S., Rosenthal, R.: Synchrony, pseudo synchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. Journal of Personality and Social Psychology 54(2), 243–253 (1988)
9. Ashenfelter, K.T., Boker, S.M., Waddell, J.R., Vitanov, N.: Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. J. Exp. Psychol. Hum. Percept. Perform. 35(4), 1072–1091 (2009)
10. Campbell, N.: Multimodal processing of discourse information; the effect of synchrony. In: 2008 Second International Symposium on Universal Communication, pp. 12–15 (2008)
11. Varni, G., Volpe, G., Camurri, A.: A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. IEEE Transactions on Multimedia 12(6), 576–590 (2010)
12. Altmann, U.: Studying movement synchrony using time series and regression models, 23 (2011)
13. Ramseyer, F., Tschacher, W.: Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. Journal of Consulting and Clinical Psychology 79(3), 284–295 (2011)

14. Delaherche, E., Chetouani, M.: Multimodal coordination: exploring relevant features and measures. In: Second International Workshop on Social Signal Processing, ACM Multimedia 2010 (2010)
15. Sun, X., Truong, K.P., Pantic, M., Nijholt, A.: Towards visual and vocal mimicry recognition in human-human interactions. In: Tunstel, E., Nahavandi, S., Stoica, A. (eds.) IEEE International Conference on Systems, Man, and Cybernetics, SMC 2011: Special Session on Social Signal Processing, pp. 367–373. IEEE Computer Society Press, USA (2011)
16. Boker, S.M., Xu, M., Rotondo, J.L., King, K.: Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. Psychological Methods 7(3), 338–355 (2002)
17. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
18. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology 18(11), 1473–1488 (2008)
19. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. 104(2), 90–126 (2006)
20. 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)
21. Mokhber, A., Achard, C., Milgram, M.: Recognition of human behavior by space-time silhouette characterization. Pattern Recognition Letters 29(1), 81–89 (2008)
22. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Proceedings of the Ninth IEEE International Conference on Computer Vision 2003, pp. 726–733 (October 2003)
23. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: First International Workshop on Spatial Coherence for Visual Motion Analysis (2004)
24. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance 2005, pp. 65–72. IEEE (October 2005)
25. Niebles, J.C., Wang, H., Fei-fei, L.: Unsupervised learning of human action categories using spatial-temporal words. In: British Machine Vision Conference (BMVC), vol. 3, p. 1249–1258 (2006)
26. Rao, C., Yilmaz, A., Shah, M.: View-Invariant Representation and Recognition of Actions (2002)
27. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 2, pp. II-123–II-130. IEEE (2001)
28. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV, pp. 432–439 (2003)
29. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady 10, 707 (1966)
30. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3), 443–453 (1970)

31. Müller, M.: Information Retrieval for Music and Motion. Springer-Verlag New York, Inc., Secaucus (2007)
32. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology 147(1), 195–197 (1981)
33. Delaherche, E., Boucenna, S., Karp, K., Michelet, S., Achard, C., Chetouani, M.: Social coordination assessment: Distinguishing between form and timing. In: Multimodal pattern recognition of social signals in human computer interaction, (submitted, 2012)

Author Index