Friedhelm Schwenker
Stefan Scherer
Louis-Philippe Morency (Eds.)

LNAI 7742

Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction

First IAPR TC3 Workshop, MPRSS 2012
Tsukuba, Japan, November 2012
Revised Selected Papers

Lecture Notes in Artificial Intelligence 7742

Subseries of Lecture Notes in Computer Science

Friedhelm Schwenker Stefan Scherer
Louis-Philippe Morency (Eds.)

Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction

First IAPR TC3 Workshop, MPRSS 2012
Tsukuba, Japan, November 11, 2012
Revised Selected Papers

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Friedhelm Schwenker
Ulm University
Institute of Neural Information Processing
89069 Ulm, Germany
E-mail: friedhelm.schwenker@uni-ulm.de

Stefan Scherer
Louis-Philippe Morency
University of Southern California
Institute for Creative Technologies
Multimodal Communication and Computation Laboratory
Playa Vista, CA 90094, USA
E-mail:{scherer, morency}@ict.usc.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-37080-9 e-ISBN 978-3-642-37081-6
DOI 10.1007/978-3-642-37081-6
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013932568

CR Subject Classification (1998): H.5, I.4, I.5, I.2.10, I.2.6, G.3, F.2.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

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

This volume presents the proceedings of the First IAPR TC3 Workshop on Pattern Recognition of Social Signals in Human–Computer Interaction (MPRSS 2012). This unique workshop endeavored to bring recent research in pattern recognition methods and human–computer interaction together, and succeeded to install a persistent forum for ongoing discussions. In recent years, research in the field of intelligent human–computer interaction has received considerable attention, and a wide range of advancements in methodology and application could be achieved. However, building intelligent artificial companions capable of interacting with humans, in the same way humans interact with each other, remains a major challenge in this field. Such interactive companions need to be capable of perceiving information about the user and its environment in order to be able to produce appropriate responses. Pattern recognition and machine learning aspects play a major role in this pioneering research. MPRSS 2012 mainly focused on pattern recognition, machine learning, and information-fusion methods with applications in social signal processing, including multimodal emotion recognition, recognition of human activities, and estimation of possible user intentions. High quality across such a diverse field can only be achieved through a selective research process. For this workshop, 21 papers were submitted out of which 13 were selected for presentation at the workshop and inclusion in this volume.

MPRSS 2012 was held as a satellite workshop of the International Conference on Pattern Recognition (ICPR 2012) held in Tsukuba, Japan, on November 11, 2012. It was supported by the University of Ulm (Germany), the University of Southern California (USA), and the Transregional Collaborative Research Center SFB/TRR 62 *Companion-Technology for Cognitive Technical Systems* at the University of Ulm and Otto von Guericke University Magdeburg, the International Association for Pattern Recognition (IAPR), and the IAPR Technical Committee on Neural Networks and Computational Intelligence (TC 3). TC 3 is one of 20 Technical Committees of the IAPR, focusing on pattern recognition applications based on methods from the fields of computational intelligence and artificial neural networks. We are grateful to all authors who submitted their manuscripts to the workshop. Special thanks to the local organization staff of the ICPR main conference for supporting MPRSS. The contribution from the members of the Program Committee in promoting MPRSS and reviewing the papers is gratefully acknowledged. Finally, we wish to express our gratitude to Springer for publishing these proceedings in their LNCS/LNAI series, and for their constant support.

December 2012

Friedhelm Schwenker
Stefan Scherer
Louis-Philippe Morency

Organization

Organizing Committee

Friedhelm Schwenker	University of Ulm, Germany
Stefan Scherer	University of Southern California, USA
Louis-Philippe Morency	University of Southern California, USA

Program Committee

Nick Campbell, Ireland
Anna Esposito, Italy
Jonghwa Kim, Germany
Bernd Michaelis, Germany
Heiko Neumann, Germany
Günther Palm, Germany
Lionel Prevost, France
Björn Schuller, Germany
Harald C. Traue, Germany
Edmondo Trentin, Italy
Michel Valster, The Netherlands
Alessandro Vinciarelli, UK

Sponsoring Institutions

Ulm University (Germany)
University of Southern California (USA)
International Association for Pattern Recognition (IAPR)
IAPR Technical Committee 3 (TC3) on *Neural Networks and Computational Intelligence*
Transregional Collaborative Research Center SFB/TRR 62 *Companion-Technology for Cognitive Technical Systems* at the University of Ulm and Otto von Guericke University Magdeburg

Table of Contents

Motion Analysis and Activity Recognition

Multimodal Fusion

Generative Modelling of Dyadic Conversations: Characterization of Pragmatic Skills During Development Age

Anna Pesarin[1], Monja Tait[5], Alessandro Vinciarelli[2,3], Cristina Segalin[1], Giovanni Bilancia[5], and Marco Cristani[1,4]

[1] University of Verona, Italy
[2] University of Glasgow, UK
[3] Idiap Research Institute, Switzerland
[4] Istituto Italiano di Tecnologia (IIT), Genova, Italy
[5] Accademia di Neuropsicologia dello Sviluppo (A.N.Svi.), Parma, Italy

Abstract. This work investigates the effect of children age on pragmatic skills, i.e. on the way children participate in conversations, in particular when it comes to turn-management (who talks when and how much) and use of silences and pauses. The proposed approach combines the extraction of "Steady Conversational Periods" - time intervals during which the structure of a conversation is stable - with Observed Influence Models, Generative Score Spaces and feature selection strategies. The experiments involve 76 children split into two age groups: "pre-School" (3-4 years) and "School" (6-8 years). The statistical approach proposed in this work predicts the group each child belongs to with precision up to 85%. Furthermore, it identifies the pragmatic skills that better account for the difference between the two groups.

Keywords: Turn-Management, Conversation Analysis, Pragmatics, Social Signal Processing.

1 Introduction

Pragmatics investigates "*how speakers organize what they want to say in accordance to who they're talking to, where, when and under what circumstances*" [18]. Hence, the development of pragmatic skills is a crucial step towards effective interaction with others for both humans [17] and artificial agents [3]. This work investigates pragmatic skills of children in developmental age and, in particular, it shows that statistical models of turn-management (who talks when and how much) and silence - two of the most important aspects of pragmatics - predict with satisfactory performance the age group of developing children. In other words, the work shows that age influences children pragmatics to an extent sufficient to be automatically detected with machine intelligence approaches.

The proposed approach extracts Steady Conversational Periods (SCP) [4] from conversation recordings and feeds them to Observed Influence Models (OIM) [15]. Then, it applies Generative Score Spaces (GSS) [13] and feature

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 1–8, 2013.

selection strategies to distinguish between models trained over conversations involving children of different age groups. The experiments were performed over a corpus of 38 conversations involving two children each (76 subjects in total). Half of the conversations include children in *pre-School* (pS) age, while the other half include children in *School* (S) age. The children of the pS group are 3-4 years old, while the others are 6-8 years old.

The results show that children can be automatically assigned to the correct age group with precision up to 85%. Furthermore, the use of GSS and feature selection shows that the pragmatic aspects that better discriminate between the two age groups are (i) the probability of observing a long silence after a long period of sustained conversation, (ii) the probability of observing short periods of sustained conversation after long silences, and (iii) the probability of observing a long silence after a short period of sustained conversation. Overall, the probabilities above suggest that S children manage to sustain conversation for longer periods and more frequently than pS children.

The rest of the paper is organized as follows: Section 2 provides a brief overview of related work, Section 3 illustrates the proposed methodology, Section 4 reports on experiments and results, and Section 5 draws some conclusions.

2 Related Work

Both development and computing literature propose a large number of works where pragmatics related measurements (e.g., total speaking time, statistics of turn length, prosody, voice quality, etc.) are shown to be the evidence of social and psychological phenomena.

From a development point of view, most of the literature focused on the interaction between gestures and first words of the child, with particular attention to the phylo-ontogenetic origin of human language and its hypothetical link with the premotor system [7]. Several researchers examined the development of skills like decoding and production of pragmatic discourse parts like, e.g., intonation and verbal prosody [8,12]. Recently, prosodic features related to voice quality have also gained some attention as effective indicators of different emotional states and attitudes of the speaker. A branch of research in fact, focuses on the evolution of conversational qualities in age of development, studying temporal features of the speech such as turns, duration, overlapping, and communication effectiveness [2].

Measurable evidences of pragmatics were extensively investigated in the computing community as well (see [16] for an extensive survey). Examples include the work in [10], where a dialogue classification system discriminates three kinds of meetings using probability transitions between periods of speech and silence, the experiments in [9], where features based on talkspurts and silence periods (e.g., the total number of speaking turns and the total speaking length) model dominance, the approach of [11], where intonation is used to detect development problems in the early childhood, and the work in [14], where prosody analysis allows the identification of language impaired children.

3 The Approach

In line with [4], the first step of the approach is the extraction of *Steady Conversation Periods* (SCP), turn management features extracted directly from audio signals: at every moment, every conversation participant i is in a state $k_i \in [0, 1]$, where 0 accounts for the participant being silent and 1 accounts for the participant speaking ($i = 1, \ldots, C$, where C is the total number of conversation participants). A SCP is the time interval between two consecutive state changes (not necessarily of the same participants). Hence, there is a sequence of SCPs for each participant i: $\{(d(n), k_i(n)\}$, where $d(n)$ is the duration of the SCP and $k_i(n)$ is the state of speaker i in SCP n. Length of the sequence and duration $d(n)$ of every sequence element are the same for all participants because the SCP changes whenever any of the participants changes state.

Overall, the extraction of the SCPs corresponds to a segmentation of the conversation into intervals during which the configuration (who talks and who is silent) is stable. In order to take into account different durations while keeping the number of states in the Observed Influence Model finite (see below), the durations $d(n)$ are grouped into two classes (*short* and *long*) by an unsupervised Gaussian clustering performed over a training dataset.

3.1 The Observed Influence Model

The Observed Influence Model (OIM) [15] is a generative model for interacting Markov chains. For a chain i ($i = 1, \ldots, C$, where C is the total number of chains), the transition probability between two consecutive states $S_i(t-1)$ and $S_i(t)$ is:

$$P(S_i(t)|S_1(t-1), \ldots, S_C(t-1)) = \sum_{j=1}^{C} {}^{(i,j)}\theta P(S_i(t)|S_j(t-1)) \tag{1}$$

where $1 \leq i, j \leq C$, ${}^{(i,j)}\theta \geq 0$, $\sum_{j=1}^{C} {}^{(i,j)}\theta = 1$, and $P(S_i(t)|S_j(t-1))$ is the probability of chain i moving to state $S_i(t)$ at step t when chain j is in state $S_j(t-1)$ at step $t-1$. An OIM can be defined as $\lambda =< A^{(i,j)}, \pi, \theta > (1 \leq i, j \leq C)$ where $A^{(i,j)}$ is the matrix such that $A_{kl}^{(i,j)} = P(S_i(t) = l|S_j(t-1) = k)$, π is a $C \times L$ (L is the total number of states) matrix such that $\pi_{ik} = P(S_i(1) = k)$ and θ is a $C \times C$ weights matrix where $\theta_{ij} = {}^{(i,j)}\theta$. In our case, we have dialogic conversations, i.e., $C = 2$; we have also $L = 4$ states since we have two classes (short, long) for each kind of SCP (silence, speech).

3.2 Generative Score Space

Generative Score Spaces (GSS) allow one to discriminate between generative models trained over samples belonging to different classes [13]. Their ultimate goal is to combine the advantages of both generative and discriminative approaches. In particular, the explanatory power of the parameters for the former

and the higher classification accuracy for the latter. Given a sequence of observations $\{O_t\}$, and a family of generative models $\{P(O_t|\lambda)\}$ the GSS maps the observations into a features vector ${\psi_F}^f$ of a fixed dimension for each data sample.

$$ {\psi_F}^f = F\left(f\left(\{P\left(O|\lambda\right)\}\right)\right), \tag{2} $$

where f is a function induced by generative models and F is some operator applied to it. In our case, where $C = 2$, $\{O_t\}$ is a sequence of SCPs that identifies a conversation, f is the function that estimates the transition probability matrices $A^{(i,j)}$ learned on $\{O_t\}$ [1]and F is the following operator:

$$ F\left(A_{kl}^{(i,j)}\right) = \frac{1}{2}\left(A_{kl}^{(i,j)} + A_{kl}^{(j,i)}\right) \text{ if } i \neq j; \quad F\left(A_{kl}^{(i,i)}\right) = \frac{1}{2}\left(A_{kl}^{(1,1)} + A_{kl}^{(2,2)}\right) \tag{3} $$

It basically considers inter and intra probability values, averaging over the different speakers, reaching thus invariance with respect to the speakers order. At the end, avoiding repeated values, the feature vector ${\psi_F}^f$ has size $2L^2$.

4 Experiments

The goal of the experiments is to investigate the effect of age on pragmatic skills for children between 4 and 8 years old. The analysis includes two main steps, the first is the quantitative analysis of silence and speech, the second is a psychological interpretation of the OIM parameters after training over *pre-School* or *School* children (see below).

4.1 The Data

The corpus used for the experiments includes 38 dyadic conversations between Italian children of the same age (76 subjects in total). The corpus is split into two parts, 19 conversations involve 3-4 years old children, named *pre-School* (pS) hereafter, for a total of 38 subjects. The other 19 conversations include 6-8 years old children, named *School* (S) hereafter, for a total of another 38 subjects. The experimental setting corresponds to a *controlled observation* (see Figure 1), the children sit close to one another and fill an album, in a situation not particularly different from their everyday experience. The average duration of the conversations is 15 minutes and 31 seconds for pS children and 15 minutes and 21 seconds for S children. The conversations have been recorded with an unobtrusive Samsung Digital Camera *34*×.

Data was manually processed independently by two different annotators, in order to perform error-free source separation; as silence periods we considered segments that don't contain sounds; sounds like cough, sneezing, ambient noise. As speech, we considered all other segments that contain verbal sounds. Silences shorter than 600 ms have been considered part of a *speech* segment.

[1] We found that considering the coefficients ${}^{(i,j)}\theta$ does not help in the classification.

Fig. 1. Experimental setting. The children sit close to one another and fill an album.

Table 1. Amount of silence and speech SCPs for each class

Class	Silence SCP	Speech SCP
pS	74%	26%
S	72%	28%

Table 2. Mean values (sec.) for short and long SCPs

Class	Short Silence	Long Silence	Short Speech	Long Speech
pS	1.48	21.89	1.41	3.84
S	1.32	17.08	1.21	4.42

4.2 Quantitative Analysis of SCPs

The overall amount of silence and speech for pS and S conversations is reported in Table 1 and shows no significant differences between the two types of conversation. However, differences emerge when speech and silence SCPs are split into *short* and *long* classes using a Gaussian clustering (see Section 3) [5]. In particular, Table 2 shows that the mean of long silence durations is significantly higher for pS children. In other words, pS children tend to interrupt their conversations for longer periods, on average.

4.3 Classification and Parameters Analysis

In order to confirm the finding above, 38 OIMs were trained over the corpus, one over each conversation. The states correspond to *short* and *long* silence and speech SCPs (four states in total). The resulting OIM parameters are mapped into a score space as described in Section 3.2 so the features extracted from each conversation are the transition probabilities between OIM states (32 features in total).

Figure 2 shows the mean values of the transition probabilities for the two types of conversation. The 5 features $F = \{f_{30}, f_{14}, f_{13}, f_{27}, f_{29}\}$ were selected as those with the highest difference between pS and S children. After this manual selection, an exhaustive feature selection procedure was applied (based on all the possible combinations of features evaluated with the K-nearest neighbor classifier [6], where K was chosen with a model selection procedure using the

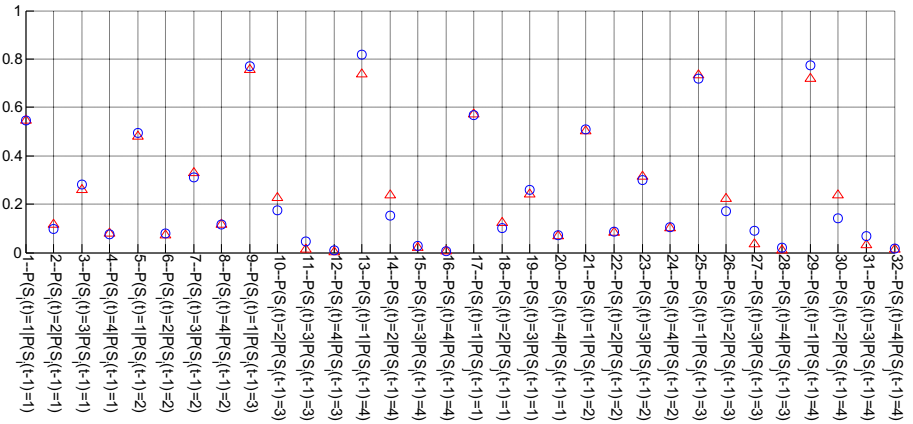


Fig. 2. Mean values of features for both classes. Triangles for pS and circles for S. Each feature has an identification number followed by its meaning.$P\left(S_i\left(t\right)=A|S_j\left(t-1\right)=B\right)$ indicate the probability of speaker S_i to be in state A after that speaker S_j was in state B at time $t-1$. The meaning of the states is: 1 = short silence, 2 = long silence, 3 = short speech, 4 = long speech.

Table 3. Classification performance of the proposed approach

Performances	pS class	S class
Precision	74%	87%
Recall	89%	68%

PRTools toolbox [1]), that led to the final feature set used for the classification experiments: $F_b = \{f_{30}, f_{27}\}$. [2]

The K-Nearest Neighbors classifier was applied using the F_b feature set as plotted in Figure 3. The classification performances, reported in Table 3, are obtained by applying a leave-one-out approach, inserting the test sequence in the training dataset, and exploiting another sequence of the pool as test.

The classification effectiveness suggests that the selected features actually characterize the two classes. The use of OIM and GSS allows one to interpret the OIM parameters under a psychological point of view. Feature f_{30} is related to the probability of transition between long speech intervals of one speaker and long silences of the other, and is higher for pS subjects.

The other feature used for the classification is f_{27}, which is an inter-speaker probability that accounts for the transition between two short speech states; its values confirm that the S conversational rhythm is higher than pS subjects.

[2] It is worth noting that each feature of F_b, taken independently, gave rise to two sets of feature values (one for each class), which were statistically independent considering the student's t-test. In particular, the null hypothesis was rejected with a significance level of 3%.

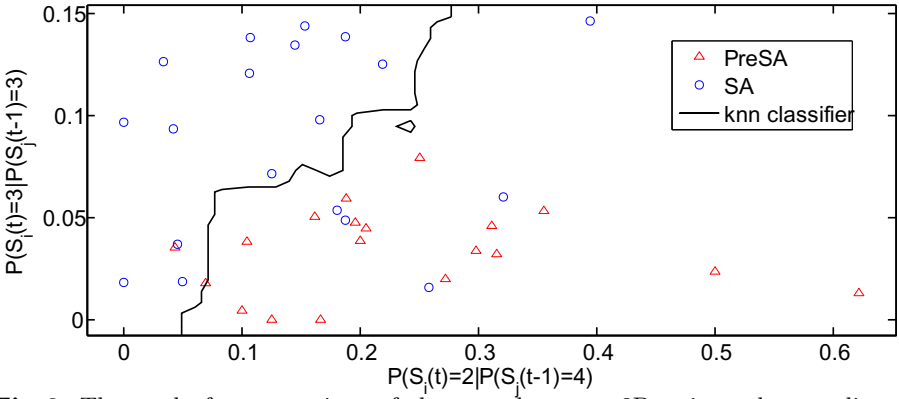


Fig. 3. The pool of conversations of the two classes as 2D points: the coordinates are the features selected by exhaustive feature selection on the set F, i.e., $f_{30} = P\left(S_i(t) = 2|P(S_j(t-1) = 4\right)$ (x-axis), $f_{27} = P\left(S_i(t) = 3|P(S_j(t-1) = 3\right)$ (y-axis)

Indeed, this transition can occur when we are in presence of overlapping speech or (less frequently) an alternation of speech periods without pauses inside.

Overall, the results showed that S subjects seem to keep a higher conversational rhythm compared to pS subjects.

5 Conclusion

This paper offers a novel study of how effectively turn taking markers can discriminate the age of children. The use of Steady Conversational Periods, fed into hybrid classifiers, allowed to finely separate classes of pre-scholar and scholar conversations, explaining actually how the two classes are different: scholar children tend to have longer and more frequent periods of sustained conversation. This study promotes many future developments: considering children of different nationalities may generalize the results obtained; more importantly, this approach may lead to the definition of a clinical semeiotics able to individuate automatically pragmatic language impairments, such those that characterize autism.

References

1. Prtools version 4.1: A matlab toolbox for pattern recognition. Internet (2004), `http://www.prtools.org`
2. Bishop, D.V., Adams, C.: Conversational characteristics of children with semantic-pragmatic disorder. ii: What features lead to a judgement of inappropriacy? The British Journal of Disorders of Communication 24(3), 241–263 (1989)
3. Cassell, J.: Embodied conversational interface agents. Communications of the ACM 43(4), 70–78 (2000)
4. Cristani, M., Pesarin, A., Drioli, C., Tavano, A., Perina, A., Murino, V.: Generative modeling and classification of dialogs by a low-level turn-taking feature. Pattern Recogn. 44(8) (2011)

5. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39, 1–38 (1977)
6. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley and Sons (2001)
7. Fogassi, L., Ferrari, P.F.: Mirror neurons and the evolution of embodied language. Current Directions in Psychological Science 16(3), 136–141 (2007)
8. Friend, M.: Developmental changes in sensitivity to vocal paralanguage. Developmental Science 3(2), 148–162 (2000)
9. Hung, H., Huang, Y., Friedl, G., Gatica-perez, D.: Estimating the dominant person in multi-party conversations using speaker diarization strategies. In: ICASSP (2008)
10. Laskowski, K.: Modeling vocal interaction for text-independent classification of conversation type. In: Proc. SIGdial, pp. 194–201 (2007)
11. Mahdhaoui, A., Chetouani, M., Cassel, R., Saint-Georges, C., Parlato, E., Laznik, M., Apicella, F., Muratori, F., Maestro, S., Cohen, D.: Computerized home video detection for motherese may help to study impaired interaction between infants who become autistic and their parents. International Journal of Methods in Psychiatric Research 20, e6–e18 (2011)
12. Morton, J.B., Trehub, S.E.: Children's understanding of emotion in speech. Child Development 72(3), 834–843 (2001)
13. Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N.: Free energy score space. Advances in Neural Information Processing Systems 22, 1428–1436 (2009)
14. Ringeval, F., Demouy, J., Szaszák, G., Chetouani, M., Robel, L., Xavier, J., Cohen, D., Plaza, M.: Automatic intonation recognition for the prosodic assessment of language impaired children. IEEE Transactions on Audio, Speech and Language Processing 19(5), 1328–1342 (2011)
15. Basu, S., Choudhury, T., Clarkson, B., Pentland, A.: Towards measuring human interactions in conversational settings. In: IEEE Int'l Workshop on Cues in Communication (CUES 2001), Hawaii, CA (2001)
16. Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A.: Social signals, their function, and automatic analysis: a survey. In: IMCI 2008: Proceedings of the 10th International Conference On Multimodal Interfaces (2008)
17. Wharton, T.: The Pragmatics of Non-Verbal Communication. Cambridge University Press (2009)
18. Yule, G.: Pragmatics. Oxford University Press (1996)

Social Coordination Assessment: Distinguishing between Shape and Timing

Emilie Delaherche, Sofiane Boucenna, Koby Karp, Stéphane Michelet, Catherine Achard, and Mohamed Chetouani

Institute of Intelligent Systems and Robotics,
University Pierre and Marie Curie, 75005 Paris, France
{delaherche,boucenna,karp,michelet}@isir.upmc.fr,
{catherine.achard,mohamed.chetouani}@upmc.fr
http://www.isir.upmc.fr/

Abstract. In this paper, we propose a new framework to assess temporal coordination (synchrony) and content coordination (behavior matching) in dyadic interaction. The synchrony module is dedicated to identify the time lag and possible rhythm between partners. The imitation module aims at assessing the distance between two gestures, based on 1-Class SVM models. These measures discriminate significantly conditions where synchrony or behavior matching occurs from conditions where these phenomenons are absent. Moreover, these measures are unsupervised and could be implemented online.

Keywords: Behavior matching, synchrony, unsupervised model.

1 Introduction

Natural conversation is often compared to a dance for the exchange of signals (prosody, gesture, gaze, posture) is reciprocal, coordinated and rhythmic. Rapport building, the smoothness of a social encounter and cooperation efficiency are closely linked to the ability to synchronize with a partner or to mimic part of his behavior. Human interaction coordination strategies, including behavior matching and synchrony are yet delicate to understand and model [1]. However, the close link between coordination and interaction quality bears promising perspectives for researchers building social interfaces, robots, and Embodied Conversational Agents [2].

Many terms related to coordination co-exist in the literature. But we usually distinguish between behavior matching [3] and synchrony. Mirroring; mimicry [4]; congruence and the chameleon effect [5] are related to behavior matching. These concepts concern non-verbal communicative behaviors, such as postures, mannerisms or facial displays, and indicate similar behaviors by both social partners; the analyzed features are isomodal and qualitative.

Synchrony is related to the adaptation of one individual to the rhythm and movements of the interaction partner [3,6,7] and the degree of congruence between the behavioral cycles of engagement and disengagement of two people.

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 9–18, 2013.

In opposition to behavior matching, synchrony is a dynamic phenomenon and can intervene across modalities.

These definitions are theoretic and in practice both forms of coordination can be observed at the same time. In this paper, we argue that despite the co-existence of both phenomenon in social interactions, a unique system is not adequate to model both forms of coordination. We propose to create two models: one dedicated to characterize synchrony and another system to assess behavior matching. In this paper, we will focus on behavior matching assessment. Synchrony assessment is described succinctly and will be detailed in future work.

2 Previous Works and Proposed Approach

Actual state-of-the-art methods to assess synchrony are based on correlation. After extracting the movement time series of the interactional 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 of the predictive association between two time series (i.e., the peak cross-correlation that is closest to a lag of zero) [8]. The main flaw of these methods is the mixing between the temporal and content aspects of coordination. Correlation informs on the temporal relation between events. But the similarity between the shape of events is poorly treated as gestures are often inadequately represented (e.g. motion energy).

In this paper, we propose to differentiate the temporal and the content part of coordination. We propose the following architecture (see Fig 1). A first module detects the onsets of gestures of both partners by identifying a strong increase of motion energy. Two modules receive the timings of the gestures : the synchrony module and the behavior matching module. The synchrony module answers the question : are the two partners in synchrony and is there an interpersonal rhythm between the two partners? Based on the timing of the segmented gestures, several metrics qualifying the respective rhythm of each partner and their interpersonal rhythm are proposed. The behavior matching module answers the question : to which extent two gestures are similar? It first identifies for each gesture of one partner the closest gestures in time from the other partner. Then, it assesses the distance between each pair of gestures. This metric is unsupervised and does not rely on predefined actions. Indeed, we are not interested in categorizing the gestures but only in comparing them.

3 Synchrony Module

This module characterizes the dynamics of activation of the dyadic partners. We are interested in the timing of events, regardless of their shape. In this paper, we focus on movement synchrony by analyzing onsets of gestures, but events could also be verbal like back-channel vocalizations for instance. Many studies in psychology underline the importance of synchrony during the interaction between a mother and a baby. For example, babies are extremely sensitive to the

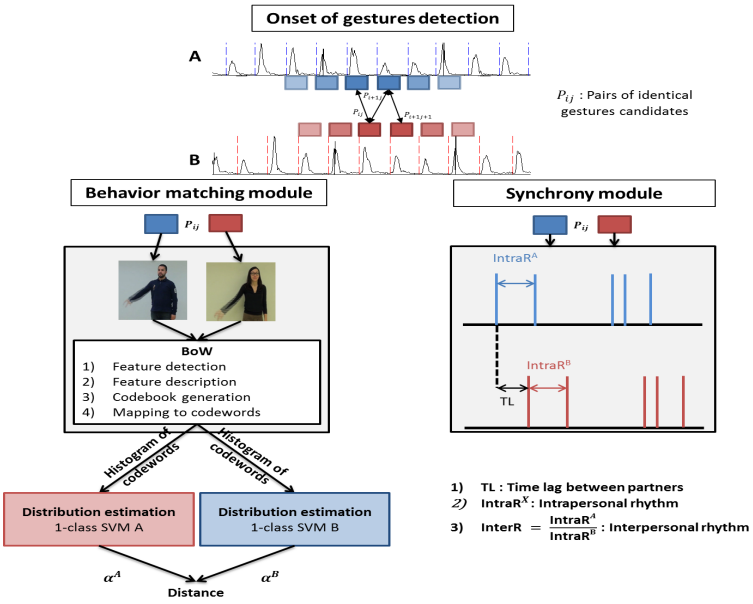


Fig. 1. A first module segments the gestures of both partners by identifying a strong increase of motion energy. Two modules receive the timings of the segmented gestures: the behavior matching module and the synchrony module.

interaction rhythm with their mother [7,6]. A social interaction rupture involves negative feelings (e.g., agitation, tears) while a rhythmic interaction involves positive feelings and smiles.

For each onset of gesture c_n^A at time $t_{c_n^A}$ detected on the partner A, the closest onset of gesture $t_{c_n^B}$ of the partner B is identified. Several features of synchrony can be extracted from these events :

- Time-lag between partners : $TL_n = t_{c_n^A} - t_{c_n^B}$ indicates which partner is leading the interaction at time c_n^A.
- Intrapersonal rhythm : $IntraR_n^A = t_{c_{n+1}^A} - t_{c_n^A}$ assesses the time between two occurrences of events for the same participant.
- Interpersonal rhythm : $InterR_n = \frac{IntraR_n^A}{IntraR_n^B}$ assesses the rapport of intrapersonal rhythm of the two partners. This measure is close to 1 if both partners share the same rhythm (whether there is a time-lag between them or not). The measure is superior to one if partner A rhythm is more important than partner's B.

The time-lag or rhythms at one moment of the interaction are not particularly informative but the variance of these features through the entire interaction, informs on whether the partners were in synchrony most of the time and adopted the same rhythm. More, in the prospect of building social interfaces, rhythm could be used as a reward signal to learn an arbitrary set of sensori-motor rules [9,10].

4 Behavior Matching Module

This module computes a distance between the dyadic partners gestures, at each time a new onset of gesture is detected. The gestures are represented with histograms of visual words and a metric based on 1-Class SVM is proposed.

4.1 Visual Features

Bag of Words models have been successfully applied in computer vision for object recognition, gesture recognition, action recognition and Content Based Image Retrieval (CBIR) [11,12,13]. The method is based on a dictionary modeling where each image contains some of the words of the dictionary. In computer vision, the words are features extracted from the image. Bag of Words models rely on 4 steps : feature detection, feature description, codebook generation, mapping to codebook. In this work, Dollàr detector [14] is used for interest point detection. It was preferred to other detectors for its robustness and for the number of interest points detected was superior, leading to a better characterization of the gesture performed. Histogram Of Oriented Gradient (HOG) and Histogram Of Oriented Flow (HOF) are used for description [12]. These descriptors characterize both shape and motion while keeping a reasonable length (compared to Dollàr descriptors for instance). The size of the feature vectors is 162 (72 bins for HOG and 90 bins for HOF). At last, it is conceivable to construct the codebook on-line with sequential k-means clustering for instance.

4.2 Distance between Gestures

We propose to derive an algorithm for novelty detection based on 1-Class SVM, proposed by Canu and Smola to estimate the distance between two gestures [15].

Distribution Estimation (1-Class SVM). 1-Class SVM was proposed to estimate the density of a unknown probability density function [16]. For $i = 1, 2, ..., n$, the training vectors h_i are assumed to be distributed according to a unknown probability density function $P(.)$. The aim of 1-class SVM is to learn from the training set a function f such that most of the data in the training set belong to the set :

$$R_h = \{h \in X \setminus f(h) \geq 0\}$$

and the region R_h is minimal. The function f is estimated such that a vector drawn from $P(.)$ is likely to fall in R_h and a vector that does not fall in R_h is not likely to be drawn from $P(.)$. The decision function is :

$$f(h) = \sum_{i=1}^{n} \alpha_i k(h, h_i) - \rho$$

The kernel $k(.,.)$ is defined over $X \times X$ by : $\forall (x_i, x_j) \in X \times X, k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_H$ where $\langle ., . \rangle_H$ denotes the dot product in H and ϕ is a mapping

from the input space X to a Reproducing Kernel Hilbert Space, called the feature space H. As in our case, h represents an histogram of codewords, we chose the histogram intersection kernel, defined as $k(h_i, h_j) = \sum_{i=1}^{d} min(h_i, h_j)$, where d denotes the size of the histogram.

Distance. Let $h_{A_i}, i = 1...n$ and $h_{B_i}, i = 1...n$ be the sequence of codewords histograms for a pair of identical gestures candidates, n denotes the size of the window. Let assume that the sequences are stationary from 1 to n and that h_{A_i} is distributed according to a distribution P_A and h_{B_i} is distributed according to a distribution P_B. To determine if the two gestures are identical, we are interested in testing the following hypothesis:

$$\begin{cases} H_0 : P_A = P_B \text{ (the gestures are identical)} \\ H_1 : P_A \neq P_B \text{ (the gestures are different)} \end{cases}$$

We write the likelihood ratio as follow :

$$L(h_{A_1}, \ldots, h_{A_n}, h_{B_1}, \ldots, h_{B_n}) = \frac{\prod_{i=1}^{n} P_A(h_{A_i}) P_B(h_{B_i})}{\prod_{i=1}^{n} P_A(h_{A_i}) P_A(h_{B_i})} = \prod_{i=1}^{n} \frac{P_B(h_{B_i})}{P_A(h_{B_i})}$$

Since both densities P_A and P_B are unknown the generalized likelihood ratio (GLR) has to be used :

$$L(h_{A_1}, \ldots, h_{A_n}, h_{B_1}, \ldots, h_{B_n}) = \prod_{i=1}^{n} \frac{\hat{P}_B(h_{B_i})}{\hat{P}_A(h_{B_i})}$$

where $\hat{P}_A$ and $\hat{P}_B$ are the maximum likelihood estimates of the densities. The exponential family gives a general representation for many of the most common distributions (normal, exponential, Poisson...). Assuming there exists a reproducing kernel Hilbert space H embedded with the dot product $< ., >_H$ and with a reproducing kernel k, the probability density function of an exponential family can be expressed :

$$P(h, \theta) = \mu(h) exp(< \theta(.), k(h, .) >_H - g(\theta)) \tag{1}$$

where $g(\theta) = log \int_X exp(< \theta(.), k(h, .) >_H) d\mu(h)$, $\mu(h)$ is the carrier density, θ is the natural parameter and $g(\theta)$ is the log-partition function. One-class SVM was proposed to estimate the support of a high dimensional distribution. Assuming that densities P_A and P_B belong to the exponential family and natural parameters θ_A and θ_B are estimated with 1-class SVM model, $\hat{P}_A$ and $\hat{P}_B$ can be written :

$$\begin{array}{l} \hat{P}_A(h) = \mu(h) exp(\sum_{i=1}^{n} \alpha_i^A k(h, h_{A_i}) - g(\theta_A)) \\ \hat{P}_B(h) = \mu(h) exp(\sum_{i=1}^{n} \alpha_i^B k(h, h_{B_i}) - g(\theta_B)) \end{array} \tag{2}$$

where α_i^A (resp. α_i^B) is determined by solving the 1-class SVM on h_{A_i} (resp. h_{B_i}). Thus,

$$L(h_{A_1}, \ldots, h_{A_n}, h_{B_1}, \ldots, h_{B_n}) = \prod_{j=1}^{n} \frac{exp(\sum_{i=1}^{n} \alpha_i^B k(h_{B_j}, h_{B_i}) - g(\theta_B))}{exp(\sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i}) - g(\theta_A))}$$

Two gestures are similar if $L(h_{A_1}, \ldots, h_{A_n}, h_{B_1}, \ldots, h_{B_n})$ is inferior to a given threshold :

$$\sum_{j=1}^{n} \left(\sum_{i=1}^{n} \alpha_i^B k(h_{B_j}, h_{B_i}) - \sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i} \right) < s_A$$

And $\sum_{i=1}^{n} \alpha_i^B k(h_{B_j}, h_{B_i})$ can be neglected in comparison with $\sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i})$. Thus two gestures are similar if :

$$\sum_{j=1}^{n} \left(- \sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i}) \right) < s_A$$

This distance can be interpreted as testing a model learned on h_{A_i} with the data from h_{B_i}. For robustness [17], we adopt the following distance in which the histograms of h_{A_i} and h_{B_i} are alternatively used for learning and for testing.

$$d = \sum_{j=1}^{n} \left(- \sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i}) \right) + \sum_{j=1}^{n} \left(- \sum_{i=1}^{n} \alpha_i^B k(h_{A_j}, h_{B_i}) \right)$$

As the visual words dictionary can be constructed online and the 1-Class SVM models are learned on the fly for each window of interaction, no supervision is required and the system can easily adapt to new gestures.

5 Results and Discussion

5.1 Data

An actual issue is the evaluation of synchrony and behavior matching models. Despite the existence of several annotating scheme, the annotation of coordination is often problematic. Indeed, the phenomenon involves the perception of complex and intricate social signals. Consequently, in several studies the measure of coordination is not validated per se, it is the ability of the measure to predict outcome variables that is evaluated.

Fig. 2. Imitation condition: the sequence of gestures

To circumvent the annotation problem, we constructed interaction data presenting different conditions of rhythm, synchrony and behavior matching. A similar approach of using simple and constructed stimuli was used to evaluate a

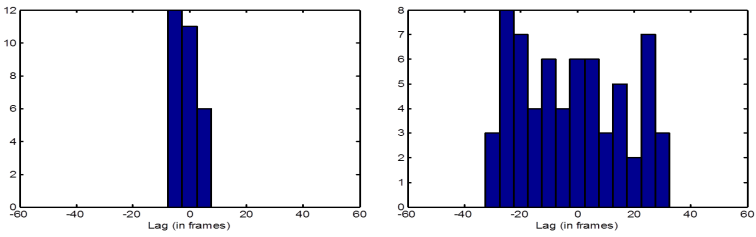
Table 1. Stimuli and conditions. We denote for each sequence its length l in seconds and the number of gestures n in the sequence l[n].

Frequency (in BPM)	Synchrony and No B.Matching (S_NBM)	Synchrony and B.Matching (S_BM)	No Synchrony and No B.Matching (NS_NBM)
20	137[44]	62[19]	
25	166[67]	71[28]	117[NA]
30	153[71]	59[27]	

model of audio-visual synchrony estimation [18]. In all conditions, two partners are standing in front of each other and filmed with a separate Sony camera at 25 fps. The focal length and focus of the cameras were optimized to capture an upper-body view of the participants. In the Synchrony (S) condition, the partners had to perform the gesture of their choice at a given rhythm, they synchronized with each other thanks to a metronome. In the Behavior Matching (BM) condition, the participants had to perform a series of identical gestures represented in Fig. 2. In total, the database contains 256 pairs of gestures including 74 pairs of identical gestures. The non-identical pairs of gestures were more numerous to account for the diversity of non-imitative situations. We voluntarily did not record a video in the NS_BM condition as it is delicate to manipulate the settings to obtain such combination. Moreover, by shifting one of the video of the S_BM condition with a certain time lag, it is possible to recreate such condition. In the NS_NBM condition, one performs at the pace of the metronome while the other is asked to gesture continually. The different conditions are summarized in Table 1.

5.2 Results

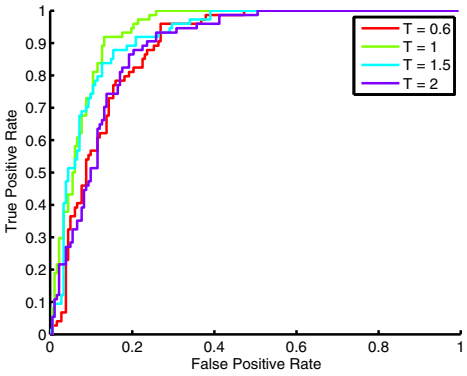
Rhythm Detection Module. To test this module, we compared the Synchrony (S) and NonSynchrony (NS) conditions. We ran this module on all the videos based on the onset of gestures identified. Figure 3 presents the histogram



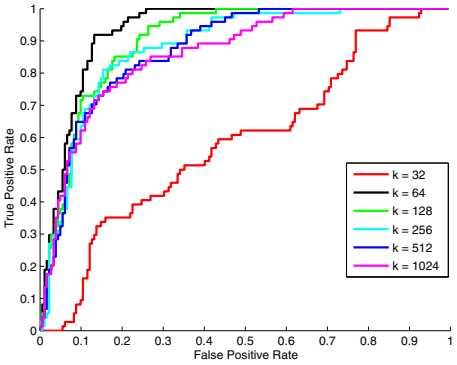
(a) Synchrony (25_S_BM) (b) Non Synchrony (25_NS_NBM)

Fig. 3. Histogram of time-lags. The variance of time-lags is larger in the NonSynchrony (NS) condition than in the Synchrony (S) condition.

of time-lags during the 25_NS_NBM and the 25_S_BM conditions. The variance of time-lag is larger in the 25_NS_NBM than in the 25_S_BM condition. We also computed the $InterR_n$ for S and NS conditions. We found that the mean and variance of $InterR_n$ were respectively 0.74 and 0.49 for the NS condition and in average the mean and variance of $InterR_n$ were respectively 0.99 and 0.13 for the S conditions. The S and NS conditions were compared with a Mann-Whitney U-test and the difference between the samples was significant (U=5147,p=7.68e-12). In the S condition, $InterR_n$ is close to 1 and varied less than in the NS condition. Moreover, $InterR_n$ is lesser than 1 in the NS condition showing that the rhythm of partner B is smaller than the rhythm of partner A. This is consistent with our scenario for the NS condition in which partner A was asked to gesture continually while partner B only gestured at the pace of the metronome.



(a) Influence of window size. k=64.



(b) Influence of codebook size. T=1s.

Fig. 4. Identical gestures classification results. The ROC curves are obtained by varying the threshold on the distance measure between pair of gestures.

Identical Gestures Detection Module. We assessed the measure of distance in the S_BM and S_NBM conditions on the segmented sequences. To test the robustness of the method, the codebook was learned on a different database than the one that serves for testing. This database was constituted with 8 videos of two different subjects performing 5 different actions composed with raising arms and waving sequences. We compared several sizes of codebook $k = 32, 64, 128, 256, 512, 1024$ and several sizes of window to assess the distance $T = 0.6, 1, 1.5$ and 2s. We performed left-tailed t-tests to compare the S_BM and S_NBM conditions. We found that the distance was significantly below in the S_BM condition compared to the S_NBM condition ($p<0.001$) for all k and T.

We finally considered a S_BM and S_NBM classification application and drew the ROC curves by varying the threshold on the distance (Fig. 4). The best results were obtained for 64 codewords and windows of 1s.The Area Under ROC curve equals 0.92. We analyzed the 23 confusions (S_NBM confused for S_BM) corresponding to the best threshold. Among them 9 corresponded to gestures in the same direction but at different levels (e.g.raising arms face /side /up), 4 to partial imitation (one arm performs the same gesture and not the other), 4 were identical gestures, 4 to completely different gestures and 2 were gestures with the same final position but with different initial positions.

5.3 Conclusion

In this paper, we proposed a new framework to assess separately synchrony and behavior matching in dyadic interactions. We proposed several metrics that discriminate efficiently synchronous from asynchronous situations and behavior matching from non-matching ones. More, assuming the codebook is created with incremental K-means, all the metrics proposed can be computed online given that no prior knowledge or training is required.

References

1. Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: A survey of evaluation methods across disciplines. IEEE Transactions on Affective Computing 3(3), 349–365 (2012)
2. Prepin, K., Pelachaud, C.: Shared understanding and synchrony emergence: Synchrony as an indice of the exchange of meaning between dialog partners. In: ICAART 2011 International Conference on Agent and Artificial Intelligence, vol. 2, pp. 25–30 (2011)
3. Bernieri, F., Rosenthal, R.: Interpersonal coordination: Behavior matching and interactional synchrony. In: Fundamentals of Nonverbal Behavior, pp. 401–432. Cambridge University Press (1991)
4. Sun, X., Nijholt, A.: Multimodal embodied mimicry in interaction. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 147–153. Springer, Heidelberg (2011)
5. Chartrand, T.L., Bargh, J.A.: The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology 76(6), 893–910 (1999)

6. Muir, D., Nadel, J.: Infant social perception. In: Slater, A. (ed.) Perceptual Development, pp. 247–285. Psychology Press (1998)
7. Murray, L., Trevarthen, C.: Emotional regulation of interactions between two-month-olds and their mothers. In: Field, T.M., Fox, N. (eds.) Social Perception in Infants, pp. 177–197. Ablex, Norwood (1985)
8. Altmann, U.: Investigation of movement synchrony using windowed cross-lagged regression. In: Esposito, A., et al. (eds.) Communication and Enactment 2010. LNCS, vol. 6800, pp. 335–345. Springer, Heidelberg (2011)
9. Andry, P., Gaussier, P., Moga, S., Banquet, J., Nadel, J.: Learning and communication in imitation: An autonomous robot perspective. IEEE Transactions on Systems, Man and Cybernetics, Part A 31(5), 431–444 (2001)
10. Boucenna, S., Gaussier, P., Andry, P.: What should be taught first: the emotional expression or the face? In: 8th International Conference on Epigenetic Robotics (EPIROB), Lucs (2008)
11. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision 73(2), 213–238 (2007)
12. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008), pp. 1–8 (June 2008)
13. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp. 524–531 (2005)
14. Dollr, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS, pp. 65–72 (2005)
15. Canu, S., Smola, A.J.: Kernel methods and the exponential family. Neurocomputing 69, 714–720 (2005)
16. Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. 13(7), 1443–1471 (2001)
17. Kadri, H., Davy, M., Rabaoui, A., Lachiri, Z., Ellouze, N.: Robust Audio Speaker Segmentation using One Class SVMs. In: Proceedings of the EURASIP EUSIPCO 2008, Suisse (2008)
18. Prince, C.G., Hollich, G.J., Helder, N.A., Mislivec, E.J., Reddy, A., Salunke, S., Memon, N.: Taking synchrony seriously: A perceptual level model of infant synchrony detection. In: Proceedings of the Fourth International Workshop on Epigenetic Robotics, pp. 89–96 (2004)

A Novel LDA and HMM-Based Technique for Emotion Recognition from Facial Expressions

Akhil Bansal, Santanu Chaudhary, and Sumantra Dutta Roy

Indian Institute of Technology, Delhi, India
akhil.engg86@gmail.com, {santanuc,sumantra}@ee.iitd.ac.in

Abstract. Over the last few years, many researchers have done a lot of work on emotion recognition from facial expressions using the techniques of image processing and computer vision. In this paper we explore the application of Latent Dirichlet Allocation, a technique conventionally used in Natural text processing, when used with Hidden Markov Model, for the same. The classification is done at an image sequence level. Each frame of an image sequence is represented by a feature vector, which is mapped to one of the words from the dictionary generated using K-means. Latent Dirichlet Allocation then models each image sequence as a set of topics. We further know the order of topics for image sequence from the order of words, which we use for classification in the next step. This is done by training a Hidden Markov Model for each emotion. The emotions dealt with are six basic emotions: happy, fear, sad, surprise, angry, disgust and contempt. We compare our results with another technique in which sequence information of words instead of topics is used by HMM for learning facial expression dynamics. The results have been presented on CK+ dataset [2]. The accuracy obtained on the proposed technique is 80.77% .The use of word-sequence in found to give better results in general.

Keywords: Emotion Recognition, Bag of Words (BoW), K-means, Latent Dirichlet Allocation (LDA), Hidden Markov Models(HMM), Topic Modeling.

1 Introduction

Over the last two decades, a lot of research is going on to automate emotion recognition from facial expressions, the emotions mostly concentrated being the six prototypic emotions (joy, surprise, anger, disgust, sadness and fear) proposed by Ekman [3]. The literature is too voluminous and diversified to be reviewed here. Bartlett et.al has done analysis using Gabor spatial filters [4].Valstar and Pantic [5] worked with a combination of GentleBoost, SVM and HMM using optical flow based motion features. Hong [6] fitted a Labelled graph to an input facial image for feature extraction , while Huang [7] used PDM for the same. Kimura and Yachida[8] applied potential net to a normalized image by applying a differential and Gaussian filter. Pantic[9] devised a rule-based system..Yang et al. [10] used dynamic Haar-like feature, while Zhao et al. [11] extended the well-known local binary feature (LBP) to the temporal domain and applied it to facial expression recognition. Cottrell [12] employed local principal component analysis (PCA). Lanitis et al. [13] interpreted face images by

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 19–26, 2013.

employing active appearance models . Lien [14] analyzed holistic face motion with the aid of wavelet-based, multi-resolution dense optical flow. Otsuka and Ohya [15] estimated facial motion in local regions surrounding the eyes and the mouth. Essa and Pentland [16] employed sophisticated 3D motion and muscle models for facial expression recognition and increased tracking stability by Kalman filtering.

In this paper, we propose a novel LDA and HMM based technique for emotion recognition from facial expressions.

Latent Dirichlet Allocation (LDA) is a generative three-level hierarchical Bayesian model, conventionally used with Bag of Words (BoW) model in text document processing, and information retrieval. In BoW model, each text document is represented by a bag of words, where each word is present in a dictionary already defined. The assumption in BoW model is that order of words doesn't in the document doesn't matter (hence called bag because in a bag there is no order in which things may be present). Now after all the documents have been represented by bags of words, LDA models words present in documents as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. The topics are hidden topics learnt in an unsupervised manner.

In the proposed approach, we use LDA on image sequences, assuming each image sequence as a document, and each frame of an image sequence represented by a word. LDA then provides an explicit representation of an image sequence as a set of topics, by assigning a topic to each frame. HMM is then used to learn the facial expression dynamics for each emotion using the sequence information of the topics in the image sequence. In the other setup, we train HMM with the sequence of words directly and compare the classification results with the proposed approach.

The rest of this paper is organized as follows. In Section 2, we describe the approach proposed. Sec. 3 is the results and discussion section. Sec. 4 summarizes the paper followed by acknowledgements and references.

2 Emotion Classification Using LDA and HMM Based Approach

The technique proposed starts with the application of BoW to an image sequence, it being analogous to a document in Natural text processing. Now, dictionary of words need to be defined. However, word in an image sequence is not off-the-shelf thing like the word in text documents. To achieve this, it includes two steps: Feature detection and Dictionary generation.

2.1 Feature Detection: Representation of Face and Extraction of Feature Vectors

To extract information for classification from the face, it is first represented by a set of MPEG-4 facial points, and some extra points as shown in Fig. 1. The points are selected around eyes, eye brows, nose and mouth, as facial expressions can be characterized by the motion-deformation information of these facial features. In all 37 facial points are used. Once the facial points have been selected in the first frame, these are tracked in the subsequent frames. For tracking any established point tracker may be used. We used Lucas Kanade tracker.

Fig. 1. Position of 37 facial points selected on face (taken from (CK+ dataset) (©Jeffrey Cohn)

Feature Vectors. Since the facial expressions can be characterized by the motion-deformation information of facial features, so we use this information only for our analysis. We extracted 9 features, which contain the displacement of various facial points selected on the face as discussed in Table 1.

Table 1. Information extracted in each feature ,the set of points used to measure corresponding feature and the direction in which displacement of facial points is measured

Feature Number and Information Contained	Set of facial points used for feature extraction	Direction of displacement measured
F1: Mouth stretch	27,28,29 and 33,34,35	Horizontal
F2:.Mouth open	30,31,32 and 24,25,37	Vertical
F3: Brow rise/lowering	2,3,5,6,12,13,15,16	Vertical
F4: Eye brow stretch	2,3,5,6 and 12,13,15,16	Horizontal
F5: Eye open	8,18 and 10,20	Vertical
F6: Displacement of outer corners of eye brows along horizontal direction	1 and 14	Horizontal
F7: Displacement of inner corners of eye brow along horizontal direction	4 and 11	Horizontal
F8: Displacement of outer corners of eye brows along vertical direction	1 and 14	Vertical
F9: Displacement of inner corners of eye brows along vertical direction	4 and 11	Vertical

We did analysis with 4 different feature vectors FV1, FV2, FV3 and FV4, such that in FV1, we measured 9 features as discussed in Table1 with each displacement measured with respect to previous frame (thus capturing local temporal information) ; in FV2, features being the same, but displacement measured with respect to neutral frame, which being the first frame for CK+ dataset(thus capturing global temporal information). FV3 was obtained by concatenating FV2 to FV1 (thus containing both local and global temporal information). In FV4, we measured the displacement of each facial point along x and y direction with respect to previous frame.

Now to account for pose variation due to rigid head motion, the displacement of each facial point is calculated relative to one of the referential points (P_9 P_{17} and P_{22} in Fig. 1). These points are so called because contractions of the facial muscles do not affect these points, and hence any displacement, if observed for these points is purely due to rigid head motion. Thus displacement of other facial points when measured relative to one of these points, cancel out their displacement due to rigid head motion.

Further, though two image sequences may show similar facial expressions, but the features may vary largely due to inter-person variations in facial features or scale variations. To solve this problem, we normalize the features. Let an image sequence I contains n frames and a p-dimensional feature vector is extracted for each frame starting from the second frame. Each normalized feature fnormalized i,j is then obtained as

$f_{\text{normalized } i,j} = f_{i,j} / f_{max}$, where $f_{max} = \max\{\text{absolute } (f_{i,j})\}$ ¥ i € {2, n} and j € {1, p}

2.2 Dictionary Generation

In feature detection step, we find a feature vector for each frame of the sequence. The next step is to convert feature vectors to words (analogous to words in text documents), for which we need a dictionary (analogous to a word dictionary). A word can be considered as a representative of several similar feature vectors. Thus words can be found using k-means clustering over all the feature vectors. Words are then defined as the centers of the learned clusters. The number of the clusters is analogous to the size of the dictionary. It may be noted that each word is a point in a feature space, and hence has the same dimensions as the feature vector. However, each word can be represented by an index in the dictionary. Now once the dictionary is defined, then to each feature vector we can assign a word, which is nearest to that feature vector. The distance measure used is using Squared Euclidean Distance.

2.3 Representation of Image Sequence as Topics

After BoW model step, each image sequence is associated with a set of words. In the next step, we apply LDA, which assigns a topic corresponding to each word in the image sequence document. And thus a document can now be associated with a set of topics. So after this step, we can think of an image sequence in terms of topics.

2.4 Expression Recognition Using Emotion Specific HMMs

LDA clusters co-occurring words into topics, and the topic probabilities provide an explicit representation of an image sequence. However for final classification of

image sequence into different emotion categories, we need a classifier. Now though LDA doesn't take into account the order in which words appear in the document, but since we find one word for each frame, we know the order of words in the image sequence document. This means that we also know the order of topics in each document. We leverage this information for classifier training in the final stage. The feature vector for the classifier is thus a set of topics, with further information contained in the order of topics. Here in this piece of research, HMM is used as classifier.

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states, which control the mixture component to be selected for each observation. A HMM model is specified by:

- The set of states, S , which are hidden
- The prior probabilities $\pi_i = P(q_i = s_i)$ which represent the probabilities of s_i being the first state of a state sequence
- The transition probabilities $P(q_{n+1} = sj \mid q_n = s_i)$ which represent the probabilities to go from state i to state j
- The emission probabilities, which characterize the likelihood of a certain observation x, if the model is in state s_i

The operation of a HMM is characterized by

- The (hidden) state sequence $Q = [q_1 q_2 \ldots\ldots q_n]$ $q_n \in S$
- The observation sequence $X = [X_1 X_2 \ldots\ldots X_{n-1}]$

A separate HMM is learnt for each emotion category, each of which can learn facial expression dynamics for that emotion category only, using the sequence of topics.

In the compared approach, we train each HMM using the sequence of words, the rest steps being the same.

2.5 Classification of a New Image Sequence

So given a new image sequence, first the facial points are selected in the first frame, and tracked in the subsequent frames using KL tracker. A feature vector is then extracted for each frame of the image sequence. Then each frame is represented as one of the words learnt during the training phase. Then for each word, one of the learnt topics can be extracted. Finally the sequence of topics is fed to each HMM trained, and the output of each HMM is compared to get the classification label.

3 Results and Discussion

Since k-means is significantly sensitive to the value of k which needs to be fixed apriori, we ran the algorithm for different values of k, and further for different values of topic counts. The experiment was done on CK+ dataset [2].We did 10-fold cross validation on entire labeled dataset. The results have been presented in Table 1 and Table 2 for proposed and compared approach respectively. We calculate 2 accuracies:

- Accuracy excluding samples which couldn't be recognized by any HMM (A1): It is calculated as P / (P + N)*100
- Accuracy including samples which couldn't be recognized by any HMM as falsely recognized samples (A2): It is calculated as P*100 / (P + N + Z).

It can be seen that some samples were not recognized by any of the HMM classifiers learnt. This happens when there are not enough variations in the training data to learn all the probabilities by HMM. This problem is expected to get solved by increasing the dataset size and including as much variations in database as possible.

Table 2. Results for Proposed Approach

Feature Vector	Word Count	Topic Count	No of Wrongly classified samples (N)	No of samples which couldn't be classi-fied (Z)	No. of Correctly Classified Samples (P)	A1	A2
1	25	25	101	0	289	74.10	74.10
2	25	25	77	1	312	80.21	80.00
3	100	25	75	0	315	80.77	80.77
4	100	25	84	0	306	78.46	78.46

The results obtained are near the state of the art results. The best accuracy for the proposed approach is 80.77 % for both A1 and A2. For the compared approach, the best accuracy achieved is 89.18% for A1 and 82.82% for A2. The results reflect that using the sequence of words for classification gives slightly better results. However then, the number of samples which couldn't be recognized by any of the HMM is normally high. The count further increases as the size of dictionary increases. This problem is almost ignorable with LDA based approach as can be inferred from results. Thus LDA based approach tends to get better when size of database is small.

Table 3. Results for Compared Approach

Feature Vector	Word Count	No of Wrongly classified samples (N)	No of samples which couldn't be classified (Z)	No. of Cor-rectly Classi-fied Samples (P)	A1	A2
1	200	41	105	244	85.61	62.56
1	50	72	17	301	80.70	77.18
2	200	34	63	293	89.60	75.13
2	50	61	18	311	83.60	79.74
3	200	33	90	267	89.00	68.46
3	25	68	4	318	82.38	81.54
4	200	33	95	272	89.18	68.00
4	50	52	15	323	86.13	82.82

4 Summary

In this paper, we explore the application of Latent Dirichlet Allocation for explicit representation of an image sequence as a set of topics. The sequence information of topics in image sequence is then used by Hidden Markov Model (HMM), for learning the dynamics of facial expression, for the classification task. It is observed that results obtained are very close to those obtained when sequence information of words is used for training of HMM. Further, LDA based approach tends to get better when size of database is small. Further, given a typical image sequence, words are more likely to be drawn from the same topic rather than different ones. Thus application of LDA helps in reducing noise which may be there till a frame is represented as a word.

The technique can potentially be used for real time implementation. This is possible since optical flow can be computed in real time on current GPUs. The technique proposed works with an image sequence and uses spatio-temporal information which is believed to contain more information than just spatial information. Further since HMM is used in the last step, hence the proposed technique can be used with image sequences of varying length, which is important as the dynamics of facial expressions may vary from person to person and also with recording device used. The simple displacement features have been used. However the beauty of the algorithm is that some other features may also be used in the first step.

The proposed technique can also be used for some other facial expression analysis, such as Action Unit recognition, with little modification. It can also be used for classifying different temporal phases of facial expression, such as neutral to peak, peak to neutral etc.

Acknowledgements. The authors gratefully acknowledge the support of the DIPR, for sponsoring the project, “Human Emotion Recognition using Computer Vision”, and providing valuable psychological inputs to the research.

References

1. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2000), Grenoble, France, pp. 46–53 (2000)
2. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. In: Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), San Francisco, USA, pp. 94–101 (2010)
3. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion, J. Personality Social Psychol. 17(2), 124–129 (1971)
4. Bartlett, G., Littlewort, M., Frank, C., Lainscsek, I., Fasel, F., Movellan, J.: Automatic recognition of facial actions in spontaneous expressions. Journal of Multimedia 1(6), 22–35 (2006)

5. Valstar, M.F., Pantic, M.: Fully Automatic Recognition of the Temporal Phases of Facial Actions. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 42(1), 28–43 (2012)
6. Hong, H., Neven, H., von der Malsburg, C.: Online Facial Expression Recognition based on Personalized Galleries. In: Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 354–359 (1998)
7. Huang, C.L., Huang, Y.M.: Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification. J. Visual Comm. and Image Representation 8(3), 278–290 (1997)
8. Kimura, S., Yachida, M.: Facial Expression Recognition and Its Degree Estimation. In: Proc. Computer Vision and Pattern Recognition, pp. 295–300 (1997)
9. Pantic, M., Rothkrantz, L.J.: Expert System for Automatic Analysis of Facial Expression. Image and Vision Computing J. 18(11), 881–905 (2000)
10. Yang, P., Liu, Q., Metaxas, D.N.: Boosting coded dynamic features for facial action units and facial expression recognition. In: CVPR, pp. 1–6 (2007)
11. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. on PAMI 29(6), 915–928
12. Padgett, C., Cottrell, G.: Representing face image for emotion classification. In: Mozer, M., Jordan, M., Petsche, T. (eds.) Advances in Neural Information Processing Systems, vol. 9, pp. 894–900. MIT Press, Cambridge
13. Lanitis, A., Taylor, C., Cootes, T.: Automatic interpretation and coding of face images using flexible models. IEEE Trans. Pattern Anal. Mach. Intell. 19(7), 743–756 (1997)
14. Lien. Automatic recognition of facial expression using hidden Markov models and estimation of expression intensity. Ph.D. Thesis, The Robotics Institute, CMU (April 1998)
15. Otsuka, T., Ohya, J.: Spotting segments displaying facial expression from image sequences using HMM. In: IEEE Proceedings of the Second International Conference on Automatic Face and Gesture Recognition (FG 1998), Nara, Japan, pp. 1998–442 (1998)
16. Essa, I., Pentland, A.: Coding, analysis, interpretation and recognition of facial expressions. IEEE

Generation of Facial Expression for Communication Using Elfoid with Projector

Maiya Hori, Hideki Takakura, Hiroki Yoshimura, and Yoshio Iwai

Graduate School of Engineering, Tottori University
101 Minami 4-chome, Koyama-cho, Tottori, 680-8550 Japan

Abstract. We propose a method for generating facial expressions with a mobile projector built into a cellphone-type tele-operated android, called Elfoid. Elfoid is designed to transmit the presence of a speaker to a communication partner in a remote place using a camera and microphone and a soft exterior that provides the look and feel of human skin. To transmit the presence of a speaker, Elfoid sends not only voice but also facial expressions and emotion information captured by the camera and microphone. Elfoid cannot, however, display facial motions because of its compactness and the lack of sufficiently small actuator motors. Therefore, we use a mobile projector and generate projection patterns to represent facial expressions estimated with a camera.

1 Introduction

Cellular phones are constantly being improved in terms of their functionality. Most of the latest cellular phones have an outstanding user interface. However, the communication function has not changed over the years and has depended on the speaker's voice. Although a video-phone can convey a speaker's facial expression, it cannot convey the human presence to a remote place.

We are conducting a collaborative project with the Advanced Telecommunications Research Institute International and Osaka University on a humanoid robot called Elfoid. The current version of Elfoid is a communication medium that downsizes Telenoid R1[1] in such a way that it can be held in the hand, and Elfoid is expected to be used instead of a cellular phone in the future. The term Elfoid is a new term coined from the word "elf" and the Latin postfix -oides, which indicates similarity, as in the word humanoid. Figure 1 shows that Elfoid has a function for communication and a soft exterior that provides a feeling of human skin. By transmitting the speaker's facial expressions and feelings information, Elfoid can convey the human presence to a remote place. To transmit facial expressions and feelings information requires robust real-time recognition of facial expressions, which are to reappear via Elfoid.

Over the past decades, there have been many studies on face recognition [2–4]. One of the latest approaches using a camera and depth sensor [4] can recognize facial expression in real time accurately. Since it is difficult to attach multiple sensors to Elfoid, a face recognition approach using video captured by a single camera [2, 3] is applied in this study. The active appearance model (AAM) [3],

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 27–34, 2013.

Fig. 1. Elfoid: cellular-phone-type teleoperated android

which is employed as part of a representative approach using video captured by a single camera, can track facial feature points in real time, although it is necessary to have training data that include shape and appearance information.

If the speaker's facial movements estimated by conventional face-recognition approaches are accurately regenerated with Elfoid, the human presence can be conveyed. However, Elfoid cannot operate like a human face because it has a compact design that cannot be activated intricately. In this research, facial expressions are generated using Elfoid's head-mounted mobile projector to overcome the problem. However, even if a captured face image is projected directly, details of facial expression cannot be conveyed because the projection plane is narrow. Few studies have investigated such a miniature device. To represent facial expressions, we generate emphasized projection patterns using the results of face recognition.

2 Generation of Facial Expression Using Elfoid with a Projector

Elfoid is used as a cellular phone for communication as shown in Fig.2. To convey the human presence, Elfoid has the following functions.

- Elfoid has a body that is easy to hold in the hand.
- Elfoid's design is recognizable at first glance to be nothing more than a human and is capable of being interpreted equally as male or female, old or young.
- Elfoid has a soft exterior that provides a feeling of human skin.
- Elfoid is equipped with a camera and microphone.

Additionally, a mobile projector is mounted in Elfoid's head and a facial expression is generated by projecting images from within the head in this study.

Fig. 2. Communication using Elfoid, which conveys the human presence to remote locations

First, individual facial images are captured by a camera mounted within Elfoid. Next, an AAM, which is generated beforehand, is used to track feature points on the face. Facial expressions are generated by warping an Elfoid image using information of the movements of feature points. Effects that induce a particular emotion are added to the image. Finally, the generated image is projected on the face of Elfoid from within.

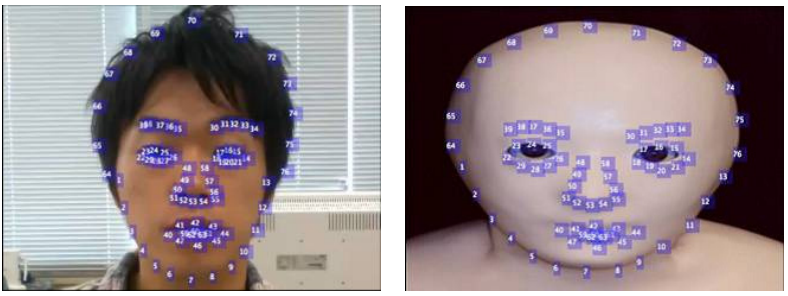
2.1 Tracking of Facial Feature Points Using an AAM

To convey facial expressions through Elfoid, face recognition using a camera mounted in Elfoid is needed. In this study, facial feature points are tracked using an AAM [3] for face recognition. The AAM is operated robustly for changes in the head pose or illumination condition, instead of requiring the generation of a model of a face in advance.

First, a facial AAM is generated using video that contains various facial expressions. A number of feature points for the eyes, eyebrows, nose, mouth, and facial outline are detected manually. Positions of the feature points are used as shape and texture information to generate a facial AAM. The AAM, which is deformed by adjusting a few parameters, is generated by applying principal component analysis (PCA) to variations of positions $s = (x_1, y_1, x_2, y_2, ..., x_n, y_n)^T$ and texture information $A(\boldsymbol{x}^T)$ at the positions $\boldsymbol{x} = (x, y)^T$. Equation (1) is a shape variation model generated by applying PCA to position information.

$$\boldsymbol{s_m} = \boldsymbol{s_0} + \sum_{i=1}^{n} p_i \boldsymbol{s}_i, \tag{1}$$

where $\boldsymbol{s_0}$ is the average of a feature's position $\boldsymbol{s}$ in a number of images. The shape vector $\boldsymbol{s_m}$ is derived from principal components $\boldsymbol{s}_i$ and weight coefficients p_i. Variations in parameter p_i can be used instead of compact principal components to produce various facial shapes.



(a) (b)

Fig. 3. (a) Feature points in the reference facial image. (b) Corresponding feature points in the facial image of Elfoid.

Equation (2) is an appearance variation model generated by applying PCA to texture information $A(\boldsymbol{x}^T)$.

$$A(\boldsymbol{x}) = A_0(\boldsymbol{x}) + \sum_{i=0}^{m} \lambda_i A_i(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \boldsymbol{s}_0, \tag{2}$$

where $A(\boldsymbol{x})$ is the appearance derived from principal components $A_i(\boldsymbol{x})$ and weight coefficients λ_i. $A_0(\boldsymbol{x})$ is the pixel values at positions $\boldsymbol{s_0}$. Variations in parameter λ_i can produce various facial appearances as in the case of $\boldsymbol{s_m}$.

Moreover, a model whose shape and appearance vary is generated by applying PCA to p_i and λ_i because the position of a feature point has high correlation with texture. The feature points are tracked robustly for changes in head pose or illumination condition by fitting the AAM to a face in an input image. The feature points can be tracked in real time using an inverse compositional algorithm [5] for the search.

2.2 Generation of Facial Expressions Using an Elfoid Image

The facial shape of Elfoid cannot be varied, the same as the case of translations of tracked feature points, because of its compactness and the lack of sufficiently small actuator motors. In this study, a facial expression is generated with Elfoid by reflecting only a variance in translation.

First, a reference facial image as shown in Fig. 3 is selected from an input video to estimate vectors of translation due to variation in facial expression. The translation vectors are estimated from positions of feature points in the reference image $\boldsymbol{P_{fb}}$ and positions of feature points in input images $\boldsymbol{P_f}$. The facial expression is generated using an Elfoid image as shown in Fig. 3. Feature points $\boldsymbol{P_{eb}}$ of the Elfoid image, which correspond to those of the reference image,

are detected manually. The Elfoid image is warped using the estimated vectors expressed as Eq. (3).

$$\boldsymbol{P_e} = \boldsymbol{P_{eb}} + \boldsymbol{W}(\boldsymbol{P_f} - \boldsymbol{P_{fb}} - \boldsymbol{T_f}), \tag{3}$$

where $\boldsymbol{T_f}$ is a translation vector of the head of the speaker, and $\boldsymbol{W}$ is a diagonal matrix that indicates a weight of translation of feature points. The weight $\boldsymbol{W}$ can vary the strength of the facial expression.

2.3 Generation of Projection Patterns for Elfoid

It is difficult to convey a facial expression accurately by only projecting the face image generated in 2.2 because Elfoid has a design that is only recognizable at first glance to be nothing but a human. In this study, projection patterns are generated to backproject Elfoid's face with consideration of Elfoid's material. According to FACS [6, 7], which describes relationships between emotion and facial movement, features around the mouth and eyebrows play important roles. In this study, movements of these feature points are emphasized by increasing the brightness of feature points or the weight $\boldsymbol{W}$ in Eq. (3).

Moreover, color stimuli that induce a particular emotion are added if a warped image can not convey a desired emotion. It is widely recognized that colors have a strong impact on our emotions and feelings [8, 9]. Facial expressions are generated by projecting feature points around the mouth and eyebrows with a particular color. When carrying out the above process, the positions of the projector and Elfoid are already calibrated.

3 Experiment

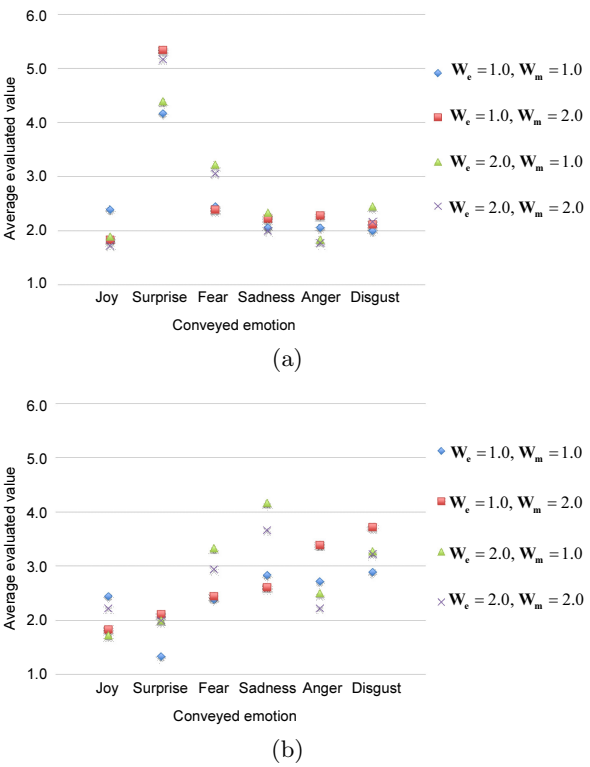
3.1 Real-Time Generation of Facial Expression Using an Elfoid Image

In an experiment, facial expressions of the communication partner are generated using an Elfoid image. First, an AAM is generated in off-line processing. The speaker's facial image is captured with a camera mounted in Elfoid, and an AAM that has 76 feature points is generated in advance. 320×240-pixel video is used to construct the AAM and track feature points. The AAM is generated from 36 facial images, which include various facial expressions.

Next, facial expressions are generated in real-time processing. The speaker's facial image is captured with a camera mounted in Elfoid, and feature points are tracked using the generated AAM. Facial expressions when the speaker shows basic emotions that are defined by Ekman et al.[7] are generated using the results of tracking. The emotions are joy, surprise, fear, sadness, anger and disgust. Figure 4 shows examples of the generated facial image when the speaker is surprised. Here, $\boldsymbol{W_e}$ is the weight for features around the eyebrows and $\boldsymbol{W_m}$ is the weight for features around the mouth. It seems that the facial expression changes depending on the weight $\boldsymbol{W}$.

(a) $\boldsymbol{W_e} = 1.0, \boldsymbol{W_m} = 1.0$ (b) $\boldsymbol{W_e} = 2.0, \boldsymbol{W_m} = 2.0$

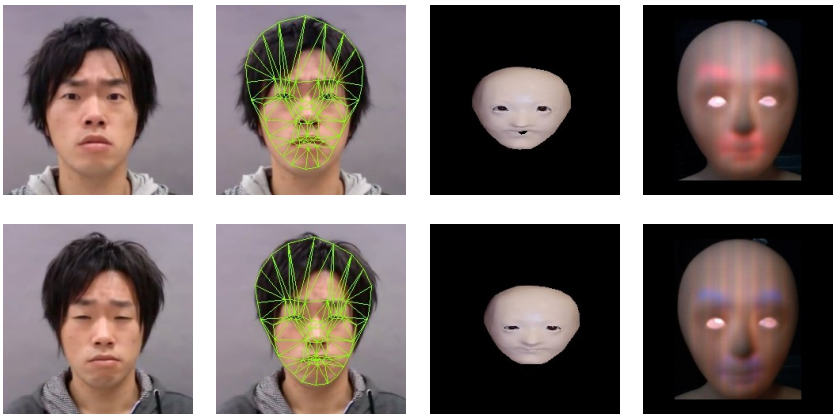
Fig. 4. Facial expressions generated by warping, when the communication partner is surprised



(a)

(b)

Fig. 5. Conveyed emotion when subjects observe the (a) surprised and (b)angry expressions



(a) (b) (c) (d)

Fig. 6. Expressions of emotion using Elfoid. The top row shows the case of anger and the bottom row shows the case of sadness. Each image shows (a) a captured image, (b) a result of feature tracking, (c) a facial expression generated by warping the image of Elfoid, and (d) a facial expression generated with Elfoid.

In this experiment, an emotion conveyed to users is investigated by subjective evaluation. Twenty-four facial patterns that are generated by varying the weight $\boldsymbol{W}$ are presented to 18 subjects in random order. Each subject rates the patterns on a scale of 1 to 6, considering impressions from the generated facial images. Figure 5 presents results of the questionnaire. As shown in Fig.5, the generated image for the emotion "surprise" conveys the target emotion adequately, such as when $\boldsymbol{W_m} = 2.0$. However, some emotions such as "anger" and "sadness" are not conveyed well, as shown in Fig.6(c), even if movements of feature points are emphasized.

3.2 Conveyance of Facial Expression with Elfoid

A facial expression cannot be conveyed even if the result of warping as shown in Fig.6 (c) is projected. Therefore, facial expressions are generated by projecting feature points around the mouth and eyebrows with a particular color. Results for backprojecting are shown in Fig. 6 (d). The top row in Fig. 6 shows the case of projecting red light for anger. The bottom row shows the case of projecting blue light for sadness. It is clear that facial expressions are conveyed well in comparison with the results of Fig.6 (c).

4 Conclusion

We proposed a method for generating facial expressions with a mobile projector built in Elfoid. In experiments, facial expressions were generated by backprojecting facial patterns to Elfoid's face. In future work, it will be necessary to

evaluate the ability of conveying facial expression. Since it is possible to direct various effects using a projector, it is likely that various motions are realizable virtually in Elfoid.

Acknowledgment. This research was supported by the JST CREST (Core Research for Evolutional Science and Technology) research promotion program "Studies on cellphone-type tele-operated androids transmitting human presence".

References

1. Ogawa, K., Nishio, S., Koda, K., Balistreri, G., Watanabe, T., Ishiguro, H.: Exploring the natural reaction of young and aged person with telenoid in a real world. Jour. of Advanced Computational Intelligence and Intelligent Informatics 15(5), 592–597 (2011)
2. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models - their training and application. Jour. of Computer Vision and Image Understanding 61(1), 38–59 (1995)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
4. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. ACM Trans. on Graphics (2011)
5. Matthews, I., Baker, S.: Active appearance models revisited. Int'l Jour. of Computer Vision 60(2), 135–164 (2004)
6. Ekman, P., Frisen, W.: Facial action coding system: A technique for the measurement of facial movement. Consulting Psychologists Press (1978)
7. Ekman, P., Frisen, W.V., Hager, J.C.: Facial action coding system (FACS). A Human Face (2002)
8. Terwogt, M.M., Hoeksma, J.B.: Colors and emotions: Preferences and combinations. Jour. of General Psychology 122(1), 5–17 (1995)
9. Kaya, N., Epps, H.H.: Relationship between color and emotion: A study of college students. College Student Journal 38, 396–406 (2004)

Eye Localization from Infrared Thermal Images

Shangfei Wang, Peijia Shen, and Zhilei Liu

Key Lab. of Computing and Communication Software of Anhui Province,
School of Computer Science and Technology,
University of Science and Technology of China, HeFei, AnHui, P.R. China, 230027
sfwang@ustc.edu.cn, {speijia,leivo}@mail.ustc.edu.cn

Abstract. By using the knowledge of facial structure and temperature distribution, this paper proposes an automatic eye localization method from infrared thermal images. A facial structure consisting of 15 subregions is proposed to extract Haar-like features. Eight classifiers are learned from the features selected by Adaboost algorithm for left and right eye, respectively. A vote strategy is used to find the most likely eyes. Experimental results on the NVIE and Equinox databases show the effectiveness of our approach.

Keywords: Eye localization, thermal infrared images, classifier, Haar-like features.

1 Introduction

Facial expressions, as the major manifestation of social signals and social behaviors [1][2], have been studied wildly in the past few years. Most researchers focus on facial expression recognition from visible images. Recently, a few researchers have paid attention to expression recognition from thermal images(IRTI), which record the temperature distribution formed by face vein branches. Since thermal images are robust to illumination variances, they are regarded as a crucial complementarity to visible images [3,4]. As eyes are one of the most important features for human face location or gaze recognition, the automatic eye localization is required for face and expression recognition in thermal spectrum. Compared with visible images, the geometric and appearance features of thermal images are more blurred. Thus, it is more difficult to locate eyes from thermal images. To the best of our knowledge, only two works [5,6] have so far been reported to detect facial components from thermal images [7]. One of them is proposed by Leonardo Trujillo et al. [5], who used Harris features and k-means clustering to detect eyes and mouth. They evaluated their approach on a gallery set composed of 30 individuals with 3 expressions (i.e. surprise, happy and angry) and 3 poses from OCTBVS database. Since the main purpose of their work is to recognize facial expression from thermal images, the detailed experimental results on eyes and mouth detection were not provided. The other work is performed by Brais Martinez et al.[6]. They adopted Haar wavelets and the GentleBoost algorithm to detect eyes and nostrils. Their approach was evaluated on their own database

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 35–42, 2013.

including 78 images of 22 subjects, and got a correct detection rate for eyes of 0.83.

Although many geometrical and appearance features are lost in thermal images, the facial structure remains, such as the eyes' symmetry. Besides, thermal images record temperature distribution on the face. For example, the eyebrow and the nose are cold [5,8], and the cheek is warm. By using this knowledge, we propose an automatic eye localization method from infrared thermal images. Our method consists of a training and a testing phase. First, the face is detected by using the thermal difference between facial area and the background. In the training phase, eight Haar-like feature sets including two edge features, four line features, one center-surrounding feature and one diagonal feature are extracted from 15 salient sub-regions around the eyes. Then the AdaBoost algorithm is used to select features from each Haar-like feature set in each sub-region. The selected features from 15 sub-regions are then combined to form a feature vector, based on which, eight classifiers using Support Vector Machines (SVM) are learned. The structure parameters of the 15 sub-regions, which are used in testing phase, are also calculated from the training samples. In the testing phase, the structure of 15 salient sub-regions is used by sliding the structure on the left/right half part of the face, and eight feature vectors are extracted. A voting strategy is used to determine whether the pixel is an eye or not. The pixel with the largest vote is declared as an eye. The proposed eye detection method is evaluated on the thermal sub-database of Natural Visible and Infrared Expression (NVIE) database [9] and Equinox database [10]. Experimental results demonstrate the effectiveness and the generalization ability of our method.

2 Method

Our approach consists of two phases, training and testing, as shown in Fig. 1. The training phase consists of face detection, feature extraction, feature selection and classification. The testing phase consists of face detection, structure sliding, feature extraction, classification and vote.

2.1 Face Detection

In order to reduce the search area of eyes, we firstly detect the face automatically. In most cases, the temperatures of human faces are different from those of the environment, so it is feasible to detect a face from thermal images. The Otsu threshold algorithm [11] is adopted to binarize infrared thermal images. Then the horizontal and vertical projection curves are calculated from binarization images. After that, the largest gradient of the projection curve is used to detect the face boundary automatically. Finally, face images are normalized to H $\times$ W, in which H and W are the height and width of face images. In order to enhance the detail of the thermal face, a histogram equalization is applied to the normalized face.

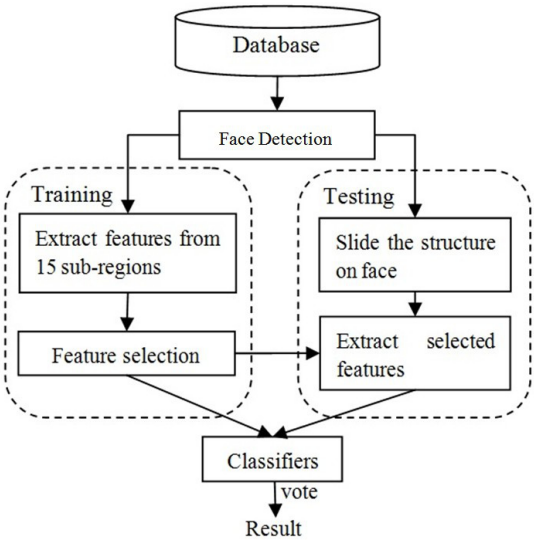


Fig. 1. Framework of our method

2.2 Feature Extraction

The geometrical and appearance features of eyes in thermal images are so weak that it is even difficult to detect eyes precisely by human beings in some cases. Therefore, it is very important to find useful characteristics in thermal images for eye location. To do this, an average thermal face is calculated from the training database, as shown in Fig. 2. From the average face, we find the temperature distributions on different facial regions are different, which is further analyzed by an Analysis of Variance (ANOVA) on the mean of sub-regions' temperature in section 3.2. For example, eyebrows and nose are the coldest part on a human face [8], the cheek is warm, and the left and right eyes are symmetric and are slightly cold. To extract useful features from these areas, we identify a structure of 15 sub-regions around them, as shown in Fig. 2. For the left eye, we assume the center of sub-region 1 is located in the left eye. Then the center of sub-region 6 is the right eye, and the center of sub-region 11 is nose. The centers of other sub-regions are determined by L and S, which are the horizontal distance between two eyes, and the vertical distance between eye and nose respectively, as shown in Fig. 2, in which, the red point indicates the center of the corresponding sub-regions. It is similar for the right eye. During the training phase, the centers of sub-regions 1, 6 and 11 are manually located, and the mean and variance of L and S can be computed from all the training samples, which will be used in the testing phase.

Then, eight kinds of Haar-like feature sets, including two edge features, four line features, one center-surrounding feature and one diagonal feature (as shown in Fig. 3) are extracted from each sub-region with size of $m \times m$ [12].

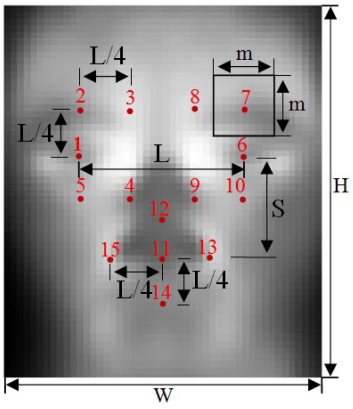


Fig. 2. The average face and 15 sub-regions

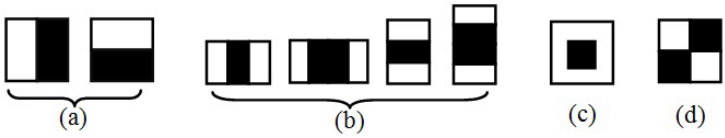


Fig. 3. Haar-like feature sets

2.3 Feature Selection and Classification

Since the feature dimension of each Haar-like set is very large, feature selection is required. The AdaBoost algorithm is used here to select features from each Haar-like feature set. After that, the selected features from 15 sub-regions are linearly combined to a feature vector with different weights, which are determined by the selected feature numbers of each sub-region. Motivated by the work of Marian Stewart Bartlett [13], SVM with linear kernel is used as the eye classifier using the Adaboost selected features. Since there are eight Haar-like feature sets, eight left eye classifiers and eight right eye classifiers are learned for left and right eye location respectively.

2.4 Testing Phase

In the testing phase, the face is detected and normalized firstly by the method described in section 2.1. When detecting the left eye, we focus on the the upper left part of the face by a sampling step of n pixels. The sampled pixels are regarded as the left eye candidates. Then eight kinds of Haar-like feature vectors are extracted from 15 sub-regions centered around the assumed left eye position, based on the structure obtained from the training phase. After that, the candidate is voted by eight well learned classifiers using the corresponding feature vectors, respectively. The pixel with the most votes is declared as the detected left eye position. A similar process is performed for the right eye localization.

3 Experiments

3.1 Experimental Condition

A set of 2067 infrared thermal frontal face images from NVIE database [9] are used as training samples, consisting of 1669 posed images and 398 first frame of spontaneous expression image sequences. During training, the centers of sub-regions 1,6 and 11 of the structure are manually labeled, thus for each image, a set of 15 sub-regions is obtained, which is used as the positive samples for localizing left and right eye in the training phase. The same amount of negative samples are randomly selected from non- eye areas with the same structure of 15 points as positive samples. These samples are used to train eight classifiers for left eye and right eye respectively.

The remaining 35,424 thermal images from NVIE database are used as the test samples to validate the effectiveness and 838 long-wave infrared images from Equinox database are used to validate the generalization capability of our method. In the training phase, we find that, after face location and normalization, L is about half of the face width for most samples, while S varies slightly. We suppose S obeys Gaussian distribution, and obtain its mean S_m and variance S_v from training samples. In the testing phase, L is set to W/2. S is set to S_m, $S_m + 2S_v$, $S_m - 2S_v$ respectively. Thus, three kinds of 15-subregion structures are used during testing. The sampling step, n, during testing phase is set to 2.

In our experiments, the width of face W is normalized to be 50, and the height is resized by the same scaling. The resolution of Haar base detector and sub-region are both 12×12, thus 11781 features, consisted of eight Haar-like features sets, are extracted from each sub-region.

3.2 Analyses of Significant Difference of Temperature among Sub-regions

We divided 15 sub-regions into 9 groups according to their locations and temperature similarity, as shown in Table 1. Then an ANOVA is performed to analyze the significant difference among the temperature mean of different groups. Based on the analysis results, we can see that the mean temperature of all group pairs except 3 pairs are significantly different at level of 0.05, as shown in Table 2.

Table 1. Groups of Facial sub-regions

Group	1	2	3	4	5	6	7	8	9
Sub-region	1, 6	2, 7	3, 8	4, 9	5,10	11	12	13,15	14

Table 2. The significant (Sig) of mean

Pairs	2 vs 4	2 vs 5	3 vs 8	other pairs
Sig	0.07	0.11	0.68	0

It demonstrates that the proposed structure captures the characteristics of the temperature variations around eyes. Thus, it may be helpful for eye localization in thermal images.

3.3 Eye Localization Results and Analyses

Motivated by the work of B. Martinez [6], we use the parameter of error, which is the displacement from automatically located centers of the target eyes to the true (manually annotated) center, to evaluate the performance of our method. The error is defined as:

$$error = \frac{\left\| P - \widehat{P} \right\|}{\|P_l - P_r\|} \tag{1}$$

where, P_l and P_r is the true position of the left and right eye, P and $\widehat{P}$ is the true and automatically detected position, respectively. $\|\cdot\|$ stands for L_2 norm.

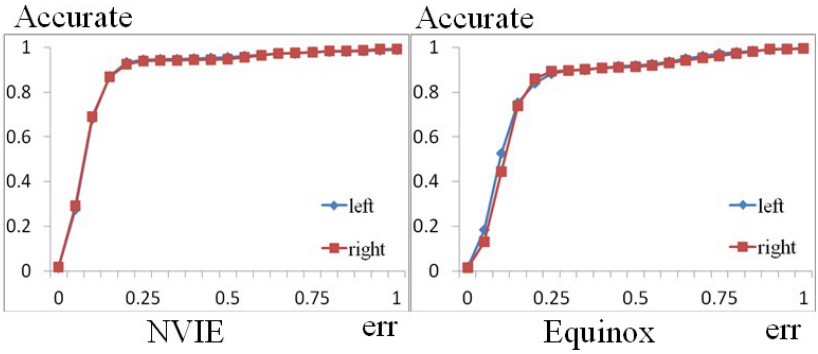


Fig. 4. Results on NVIE and Equinox

Fig. 4 shows the performance of our method. From Fig. 4 we can see that when $err < 0.15$ is regarded as success, we achieve accurate rate of localization around 88% and 75% on NVIE and Equinox database respectively. All the experiment results verify the effectiveness and acceptable generalization ability of our method. Since the classifiers are trained on the NVIE database, it is reasonable and acceptable that the accurate rate on the Equinox database is lower than that of the NVIE database. Some examples of results corresponding to the error accepted ($err < 0.15$) are shown in Fig. 5. Compared with the results of 83% achieved by B. Martinez using leave-one-subject-out cross validation in their database [6], our results are pretty competitive. The encouraging performances on the NVIE and the Equinox database demonstrate that our method is effective and robust to the changes of facial expressions, since both databases include thermal images with facial expressions.

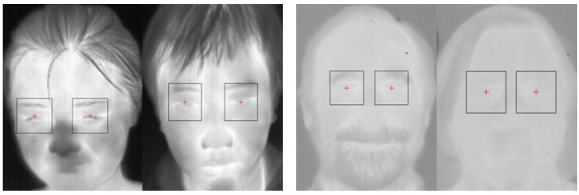


Fig. 5. Samples of the Eyes Location Results

4 Conclusion

In this paper, we aim to locate eyes in infrared thermal images. A structure consisting of by 15 sub-regions is proposed to extract the Haar-like features to capture the temperature distributions of the eyes and their adjacent facial regions. Eight classifiers are learned from the combination features selected by Adaboost algorithm for left and right eye, respectively. A vote strategy is used to find the most likely eyes. The results of ANOVA demonstrate that our structure captures the useful characteristics of the temperature distributions around eye. The eye detection experiments performed on NVIE and Equinox database verify the effectiveness and generalization ability on multi-expression infrared thermal samples.

Compared with the related two works, our contributions are as the follows: (1) Since infrared thermal images reflect the temperature distribution of human faces, we propose a 15 sub-regions structure to capture both the temperature distribution of the eyes and that of the adjacent regions for robust eye localization. (2) We evaluate our approach on the sub-database of NVIE database, including 35,424 images for 76 subjects, which is much larger than previous two research. Furthermore, We are the first to evaluate eye detection from thermal images by a cross-corpus experiment. It demonstrates the generalization ability of our approach.

Although our proposed method is comparable with previous related research, some additional works are necessary to improve the operating speed and to meet the needs of real-time applications. These will be conducted in the future.

Acknowledgements. This paper is supported by the NSFC (61175037), Special Innovation Project on Speech of Anhui Province (11010202192), project from Anhui Science and Technology Agency (1106c0805008) and Youth Creative Project of USTC.

References

1. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vision Computing 27(12), 1743–1759 (2009)
2. Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A.: Social Signal Processing: State-of-the-art and future perspectives of an emerging domain. In: Proceedings of the ACM International Conference on Multimedia, pp. 1061–1070 (2008)

3. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1), 39–58 (2009)
4. Khan, M.M., Ward, R.D., Ingleby, M.: Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature. ACM Transactions on Applied Perception (TAP) 6(1) (2009)
5. Trujillo, L., Olague, G., Hammoud, R., Hernan-dez, B.: Automatic feature localization in thermal images for facial expression recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, p. 14 (2005)
6. Martinez, B., Binefa, X., Pantic, M.: Facial component detection in thermal imagery. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), pp. 48–54 (2010)
7. Hansen, D.W., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3), 478–500 (2010)
8. Shastri, D., Pavlidis, I.: Automatic initiation of theperiorbital signal extraction in thermal imagery. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2009), pp. 182–187 (2009)
9. Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Chen, F., Wang, X.: A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Transactions on Multimedia 12(7), 682–691 (2010)
10. Equinox corporation - human identification at a distance databas, `http://www.equinoxsensors.com/products/HID.html`
11. Otsu, N.: A threshold selection method from gray-level histograms. Automatica 11, 285–296 (1975)
12. Mita, T., Kaneko, T., Hori, O.: Joint haar-like features for face detection. In: IEEE International Conference on Computer Vision (ICCV 2005), vol. 2, pp. 1619–1626 (2005)
13. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), June 20-25, vol. 2, pp. 568–573 (2005)

The Effect of Fuzzy Training Targets on Voice Quality Classification

Stefan Scherer[1,3], John Kane[2], Christer Gobl[2], and Friedhelm Schwenker[3]

[1] Institute of Creative Technologies, University of Southern California, United States
[2] Phonetics and Speech Laboratory, Trinity College Dublin, Ireland
[3] Institute of Neural Information Processing, Ulm University, Germany

Abstract. The dynamic use of voice qualities in spoken language can reveal useful information on a speaker's attitude, mood and affective states. This information may be desirable for a range of speech technology applications. However, annotation of voice quality may frequently be inconsistent across raters. But whom should one trust or is the truth somewhere in between? The current study looks first to describe a voice quality feature set that is suitable for differentiating voice qualities on a tense to breathy dimension. These features are used as inputs to a fuzzy-input fuzzy-output support vector machine (F^2SVM) algorithm, to automatically classify the voice qualities. The F^2SVM is compared to standard approaches and shows promising results. Performances for cross validation, leave one speaker out, and cross corpus experiments of around 90% are achieved.

1 Introduction

The term voice quality (henceforth VQ) refers to the timbre or coloring of a speaker's voice. For an individual speaker their VQ is composed of longer term settings of the vocal system combined with dynamic shifts in the system for communicative purposes [1]. A speaker's VQ is an important feature of paralinguistic signaling in speech and can provide the listener with information pertaining to the speaker's affective state [2]. For instance, breathy voice has been generally observed in association with intimacy and familiarity [1]. Tense voice on the other hand has been reported in more active affective states, e.g., anger and happiness [3].

It has been widely observed that VQ can provide useful insights into the intentions and mood of the speaker, and indeed VQ features have also been utilized in order to improve emotion classification [4]. It follows that robust characterization of voice qualities may be desirable for both input (i.e. recognition) and output (i.e. synthesis) ends of speech applications.

The purpose of this study is to put forward a framework for identifying voice qualities on a tense to breathy continuum. Few studies have focused on automatic classification of voice qualities using combinations of features. The main work in this area has been done in the domain of pathological voice types [5]. Hidden Markov models (HMMs) and a regression approach were employed to

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 43–51, 2013.

categorize speech signals, that were generally of a longer duration than the signals in this study. The task was to match the annotated degree (form 0 to 4) on three VQ scales, namely breathiness, roughness and deviance. Accuracies of about 50% within each of the three scales could be achieved in the study. However, the speech material used was mainly pathological voices which weakens its comparability with the present study. In this study we investigate fuzzy-input fuzzy-output support vector machine (F^2SVM) introduced in [6] for the task at hand and compare their performance to standard approaches, that do not make use of the fuzzy membership assignments provided by human experts.

The remainder of the paper is organized as follows: In Sec. 2 the utilized VQ features for the classification experiments are introduced. Along with the introduction of the speech dataset used, Sec. 3 introduces the annotations by experts, which are later used as training targets for the fuzzy classification experiments. Section 4 then briefly introduces the utilized F^2SVM, which compete against two standard non-fuzzy approaches. In Sec. 5 the results for the experiments are reported and discussed in Sec. 6. Finally, Sect. 7 concludes the paper and provides an outlook.

2 Voice Quality Features

The VQ features used in the current study were selected on the basis of being stated to be able to characterize voice qualities across the breathy to tense dimension. The features described in Sects. 2.1 - 2.5 describe aspects of the glottal source signal, which is derived using automatic inverse filtering. This is done using the pitch synchronous automatic inverse filtering (PSIAIF) method described in [7], with f_0 extracted using ESPS/*waves+* software package. The features described in Sects. 2.1 to 2.5 can then be measured on the output signal from this method. However, as the output of this method can sometimes contain uncancelled formant oscillations, which can negatively impact the features, we use one further feature which is measured without the use of inverse filtering (see Sec. 2.6).

2.1 Time Based LF Model Parameters (*Ra*,*Rk*,*Rg*,*EE*)

The most commonly used glottal source model is the Liljencrants-Fant (LF) model [8]. It is a five parameter (including f_0) model of differentiated glottal flow (i.e. the residual signal after inverse filtering if lip radiation has not been compensated for). The model has two components. The first component, the open phase, is a sinusoid function that increases exponentially and the second component is an exponential function which models the return phase.

$$U'_g(t) = \begin{cases} E_0 e^{\alpha t} sin\omega_g t & for \;\; t_o \leq t \leq t_e \\ \frac{-EE}{\epsilon T_a}(e^{-\epsilon(t-t_e)} - e^{-\epsilon T_b}) & for \;\; t_e < t < t_c \\ 0 & for \;\; t_c \leq t \leq T_0 \end{cases} \quad (1)$$

The model is generated using the time-points shown in Fig. 1, along with the parameters E_0, α and ϵ which are solved implicitly to ensure area balance above and below the zero-line (see [8] for full details of the model). The model can be fit to an inverse filtered speech signal in the time domain using the method described in [9]. From the given model configuration, one can obtain four parameters: the amplitude parameter EE (shown in Fig. 1) and three shape parameters; Rg, Rk and Ra (see Eqs. 2). These parameters have been shown to be suitable for characterizing a range of voice qualities including breathiness and tenseness [10]

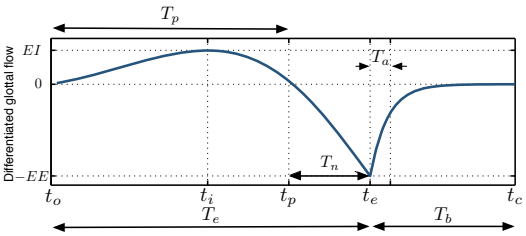


Fig. 1. Example LF model pulse for the glottal flow (above) and the differentiated glottal flow (below)

$$Rg = \frac{1}{2T_p \cdot f_0}; \quad Rk = \frac{T_e - T_p}{T_p}; \quad Ra = T_a \cdot f_0 \tag{2}$$

2.2 LF Parameters Frequency Domain (Ra_f,Rk_f,Rg_f,EE_f)

An alternative approach for deriving LF model parameters in the frequency domain was initially described in [11] and has since been further developed. The method involves using the amplitudes of the first eight harmonics from the glottal source spectrum as inputs to a feed forward neural network, previously trained on a large volume of LF model configurations and their spectral information, in order to derive the four parameters stated above. Harmonic amplitudes are measured from the narrowband spectrum, obtained by taking a three pulse length segment of the glottal source signal, centered on a GCI, and windowed using a Hamming window. This approach was developed in order to improve the robustness of the extracted parameters to the presence of noise and phase distortion.

2.3 Normalized Amplitude Quotient (NAQ)

The normalized amplitude quotient (NAQ) parameter was introduced as a global glottal source parameter capable of differentiating breathy to tense voice qualities [12]. NAQ was shown to be more robust to noise disturbances than time based parameters and has, as a result, been used widely in applied work on VQ.

2.4 $\Delta H_{1,2}$

The difference in amplitude levels (in dB) between the first two harmonics of the narrowband glottal source spectrum ($\Delta H_{1,2}$) is thought to be a rough correlate of

the open quotient parameter and hence useful at discriminating breathy to tense voice qualities [13]. The narrowband spectrum is obtained by using three-pulse length sections, centered on a GCI and using a Hamming window.

2.5 Voice Quality Spectral Gradients (OQG, GOG, SKG, RCG)

Lugger and Yang [14] described a set of spectral gradient parameters for characterizing voice qualities from glottal source signals. The parameters, comprising Open Quotient Gradient (OQG), Glottal Opening Gradient (GOG), Skewness Gradient (SKG), and Rate of Closure Gradient (RCG), were stated by the authors to be strongly correlated with typical glottal pulse shape parameters. They have been shown to be useful in the classification of voice qualities, gender and emotion, as well as relatively robust [14].

2.6 PeakSlope

A final feature is included which has recently been shown [15] to be able to separate breathy to tense voice qualities without the use of inverse filtering.

$$g(t) = -\cos(2\pi f_n t) \cdot \exp(-\frac{t^2}{2\tau^2}) \tag{3}$$

The speech segment $s(t)$ is convolved with $g(\frac{t}{s_i})$, where $s_i = 2^i$ and $i = 0,1,2,....,5$. This essentially is the application of an octave-band filter bank with the centre frequencies being: 8 kHz, 4 kHz, 2 kHz, 1 kHz, 500 Hz and 250 Hz. Then the local maximum is measured at each of the signals and a regression line is fit to these peaks and the extracted parameter is simply the slope coefficient of this regression line [15].

3 Speech Data

There is a distinctive lack of available speech data with VQ annotation. Further, as VQ annotation schemes differ and as the annotator's interpretation of VQ labels may not be consistent, this makes large scale data collection difficult. The speech data for this study comes from the recordings used in [13]. The original data were speech recordings of 6 female and 5 male speakers aged between 18 and 48 years (with a mean of 30). The speakers were asked to produce eight Finnish vowels /ɑ e i o u y æ ø/ using breathy, normal and tense phonation types. Participants were trained with producing the voice qualities before recording. While conducting the recording speakers were asked to repeat the utterance with stronger emphasis on the VQ when it was necessary. Each utterance was repeated three times resulting in 792 speech segments.

The speech was recorded using a unidirectional Sennheiser electret microphone with a preamp (LD MPA10e Dual Channel Microphone Preamplifier) and a digital audio recorder (iRiver iHP-140). Audio was digitized at 44.1 kHz.

In order to describe three independent sets of voice qualities we carried out listening test with three expert judges. All participants were experienced in VQ

research and were also familiar with Laver's labeling framework [1]. The participants rated the speech samples on a five point Likert scale from breathy (1) to tense (5). Samples were presented to the participants in a randomized order, with an inter-rater agreement of $\kappa = 0.526$. For the present study we excluded all recordings for which the maximal membership assignment did not coincide with the intended class. 478 vowel recordings were left for analysis (with an inter-rater agreement $\kappa = 0.717$).

Also, included in the current study were 10 sonorant-only (all voiced) sentences, produced in three voice qualities (breathy, modal and tense) by one male speaker (i.e. 30 sentences in total). The utterances were produced in a semi-anechoic room and audio was captured using a B&K 4191 free-field microphone and a B&K 7749 pre-amplifier.

4 Fuzzy-Input Fuzzy-Output Support Vector Machines

Support vector machines (SVM) have become one of the most popular classifiers in many different machine learning or pattern recognition applications [16]. Extended architectures like one-against-one SVM, one-against-all SVMs or tree structured SVM [17] have been developed for the classification of crisp or hard labeled data in the more recent past.

While dealing with naturalistic data, like voice qualities or user states in natural recordings, however, labels or categories might not be clear or crisp at all, but rather subjective to the perception of the annotator. Since the ground truth or the correct class might be unknown or fuzzy, the so called fuzzy SVMs (FSVM) assigning memberships to several classes to single observations have been developed by [18]. Though, the output of those FSVMs is still crisp and no fuzzy output is generated. Therefore, so called fuzzy-input fuzzy-output SVMs (F^2SVM) capable of receiving soft labeled data and producing soft outputs with memberships assigned over multiple classes have been developed [6]. The fuzzy output of the F^2SVM is required, as in the case of a multi-class one-against-one SVM (three classes in the present study) a fuzzy output is required for the proper combination of the decisions of the single SVM. Consider, for instance, that all three one-against-one SVM (i.e. in this study: breathy vs. modal; tense vs. modal; breathy vs. tense) would have different crisp opinions. Then, it would not be possible to find a sound solution for the given input. If, however, the output were fuzzy such a stalemate is unlikely.

5 Experiments and Results

In the following we have listed the results of the recognition experiments that we conducted. The standard methods of choice for comparison were naive Bayes classifier (NB), giving a rough baseline, and standard crisp SVM utilizing the same radial basis function (RBF) kernel as the F^2SVM. The approaches were compared using a standard ten fold cross validation (X-VAL; 90% training /10% test data split) as well as leave one speaker out (LOSO; for each fold one of the

Table 1. Error (in %) comparison of **NB, standard SVM and crisp F^{2}SVM** outputs for **X-VAL** and **LOSO** experiments. The error (Err.) and standard deviation (Std.) are calculated. Significant results are marked with * or **.

	X-VAL		**LOSO**	
	Err. (%)	**Std.**	**Err. (%)**	**Std.**
NB	21.54**	6.58	23.94**	10.35
SVM	16.09*	4.59	18.33*	6.99
F^{2}SVM	**12.14**	**3.11**	**13.88**	**3.89**

eleven speakers was left out of the training set and was solely used for testing) paradigms. Additionally, the generalization ability of all three methods, i.e. NB, SVM, and F^{2}SVM, is compared in a cross corpus experiment using the sentence dataset (see Sec. 3).

For the F^{2}SVM experiments it was necessary to generate fuzzy targets resembling the degree of membership of each sample towards all of the three classes. For each of the recordings these membership values were calculated using the labels (i.e. five point Likert scale), as indicated by all the experts. These newly calculated values were then used as the target signal for the F^{2}SVM in the experiments. If no clear VQ was perceived by the annotator (i.e. mixed labels 2 and 4) the same amount of membership was assigned to both voice qualities. After normalization to the number of annotators the sum of all memberships of each sample adds up to 1.

In Tab. 1 the error rates of all of the crisp classification experiments are listed. The F^{2}SVM outperforms the other baseline approaches in all experiments significantly. For the X-VAL experiments using all the available speakers 12.14% error (standard deviation $\sigma = 3.11$) was achieved, and only a slight decrease was observed while leaving one speaker out (13.88% error; $\sigma = 3.89$). In contrast to these results the standard SVM receiving the actual label as target in training resulted in 16.09% error ($\sigma = 4.59$) in the X-VAL and 18.33% ($\sigma = 6.99$) in LOSO. Both times the F^{2}SVM outperforms the standard SVM statistically significant in paired t-tests (X-VAL $p = 0.02$; LOSO $p = 0.04$). The baseline performance of the NB results in errors slightly over 20% for both the X-VAL and the LOSO experiment. Both times the NB is strongly outperformed by the F^{2}SVM with significant differences (X-VAL $p < 0.001$; LOSO $p = 0.008$). No statistically significant difference between the standard SVM and the NB was found.

The confusion matrices of these experiments can be seen in Tab. 2 (X-VAL experiment and LOSO experiment). All approaches result in very similar confusion matrices where almost no confusion between breathy and tense voice qualities are present. For the F^{2}SVM and the NB these errors are not reported in the X-VAL experiments, further, in the LOSO experiment they do not exceed 1%. In the standard SVM case breathy is confused with tense in 6% of the cases for the LOSO experiment (only 3% in the X-VAL experiment). The errors of the NB between neighboring voice qualities are, however, more frequent as in the other approaches.

Table 2. Comparison of confusion matrices using **NB, standard SVM and F^{2}SVM** approaches for **X-VAL** and for **LOSO** experiments with **all speakers** (eleven speakers). Numbers are hit rates and lines sum up to one for each confusion matrix modulo rounding errors.

		NB			SVM			F^{2}SVM		
		Breathy	**Modal**	**Tense**	**Breathy**	**Modal**	**Tense**	**Breathy**	**Modal**	**Tense**
X-VAL	**Breathy**	0.87	0.13	0.00	0.89	0.10	0.01	**0.90**	0.10	0.00
	Modal	0.19	0.65	0.16	0.13	0.78	0.09	0.08	**0.85**	0.06
	Tense	0.01	0.14	0.85	0.03	0.12	0.85	0.00	0.12	**0.88**
LOSO	**Breathy**	0.86	0.14	0.00	0.85	0.13	0.02	**0.88**	0.11	0.01
	Modal	0.20	0.62	0.18	0.13	0.78	0.09	0.09	**0.83**	0.08
	Tense	0.01	0.14	0.84	0.06	0.13	0.81	0.01	0.11	**0.88**

Table 3. Error (in %) comparison of **NB, standard SVM and F^{2}SVM** outputs for **cross corpus** experiments with frame-wise error rates as well as temporally integrated errors over full sentence length. The classifiers are trained on the Finnish vowel set and tested on the sentence data (compare Sec. 3). The error is calculated by comparing to the true label.

	Frame-wise	**Temporally integrated**
NB	29.53	30.00
SVM	33.33	30.00
F^{2}SVM	**17.66**	**3.33**

In order to further check the generalization ability of the approach a cross corpus experiment was conducted. All the mentioned methods, i.e. NB, standard SVM, and F^{2}SVM, were trained on the Finnish vowel set data and tested on the sentence dataset. The errors in % are listed in Tab. 3 comprising the errors on a frame-wise basis including vowels and consonants and the errors achieved after integrating the decisions of the approaches over the whole sentences, which were recorded in a constant VQ. It is seen, that the F^{2}SVM approach (frame-wise error 17.66%; sentence level 3.33%) again outperforms the other two reference approaches clearly. The two perform around 30% error for all cases. In the case of the sentence level integration of the decision the F^{2}SVM only mistakes one breathy sentence as a modal sentence.

6 Discussion of Statistical Evaluation

The most striking result from the experiments is the capability of the F^{2}SVM to classify the voice qualities more accurately than a standard SVM with the same features as input and kernel function (RBF kernel), in the classification experiments shown in Tab. 1. Therefore, it seems quite obvious that there is relevant information present in the fuzzy targets during training that improves the generalization capabilities of the classifier. As these experiments were conducted on the reduced dataset with an inter-rater agreement of $\kappa = 0.717$ the training of all approaches was conducted on a set for which the maximum of the annotators' membership assignments always coincides with the actual target label, in

order to render a fair comparison. Furthermore, the underlying model employed during expert annotation, described in Sec. 3, allowing the annotator to assign a label between breathy and modal (the value 2 in the Likert scale) and a value between modal and tense (the value 4 in the Likert scale) seems proven by the classification results shown in Sec. 5. This conclusion can be drawn since all the classifiers, comprising NB, standard SVM, and F^{2}SVM, confuse neighboring classes more often than the two extreme classes, breathy and tense.

Overall, the approach is sufficiently stable over untrained speakers and generalizes well. This, however, is not only the case for the fuzzy approach but also for the two baseline approaches, indicating that the features are representing the voice qualities quite well and are quite independent of the speakers (compare leave one speaker out results in Tab. 1).

The generalization capabilities of the approaches were further compared in a cross corpus experiment. The classifiers were trained using the features extracted from the Finnish vowel set data and tested on the features of the sentence data, including features corresponding to voiced-consonants and vowels alike. The F^{2}SVM clearly outperformed the reference approaches, with an accuracy of around 82% for the frame-wise decisions. Further, after integrating the decisions over the whole sentences the accuracy rose to more than 95%, meaning that one of the thirty sentences was confused.

7 Conclusion

In the present study we investigated the capability of F^{2}SVM to classify VQ samples from a vowel corpus, as well as in a cross corpus study using data taken from full sentences. The results in Sec. 5 show high accuracy rates including cross validation and leave one speaker out validation conditions. Additionally, we have shown strong generalization capabilities in cross corpus analysis and leave one speaker out experiments. The proposed method outperformed its competitors (standard SVM, and NB) in crisp classification experiments clearly, by only utilizing the information present in fuzzy labels during training. This is a very encouraging result supporting the value of fuzzy interpretations of VQ data and annotations. The results are very promising for future work including the extension of the approach to running speech and more naturalistic data.

One of the shortcomings of the present study is, that we only considered acted VQ samples. However, we believe the findings here help pave the way to improved VQ analysis in realistic speech data. The analysis of the sentence corpus is a first step into that direction and it seemingly worked very well.

Acknowledgements. The presented work was developed within the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG). The second and third authors are supported by the Science Foundation Ireland, Grant 07/CE/I1142 (Centre for Next Generation Localisation, www.cngl.ie) and Grant 09/IN.1/I2631 (FASTNET).

References

1. Laver, J.: The Phonetic Description of Voice Quality. Cambridge University Press (1980)
2. Gobl, C.: The voice source in speech communication. Ph. D. Thesis, KTH Speech Music and Hearing, Stockholm (2003)
3. Gobl, C., Ní Chasaide, A.: The role of voice quality in communicating emotion, mood and attitude. Speech Comm. 40, 189–212 (2003)
4. Lugger, M., Yang, B.: Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. In: Proc. of ICASSP, pp. 4945–4948 (2008)
5. Wester, M.: Automatic classification of voice quality: Comparing regression models and hidden Markov models. In: Proc. of VOICEDATA 1998, pp. 92–97 (1998)
6. Thiel, C., Scherer, S., Schwenker, F.: Fuzzy-Input Fuzzy-Output One-Against-All Support Vector Machines. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part III. LNCS (LNAI), vol. 4694, pp. 156–165. Springer, Heidelberg (2007)
7. Alku, P., Bäckström, T., Vilkman, E.: Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Comm. 11(2-3), 109–118 (1992)
8. Fant, G., Liljencrants, J., Lin, Q.: A four parameter model of glottal flow. In: KTH, QPSR, vol. 4, pp. 1–13 (1985)
9. Strik, H., Cranen, B., Boves, L.: Fitting a LF-model to inverse filter signals. In: Proc. of Eurospeech (ISCA), pp. 103–106 (1993)
10. Gobl, C.: A preliminary study of acoustic voice quality correlates. In: KTH, QPSR, vol. 4, pp. 9–21 (1989)
11. Kane, J., Kane, M., Gobl, C.: A spectral LF model based approach to voice source parameterisation. In: Proc. of Interspeech (ISCA), pp. 2606–2609 (2010)
12. Alku, P., Bäckström, T., Vilkman, E.: Normalized amplitude quotient for parameterization of the glottal flow. J. Acoust. Soc. Am. 112(2), 701–710 (2002)
13. Airas, M., Alku, P.: Comparison of multiple voice source parameters in different phonation types. In: Proc. of Interspeech (ISCA), pp. 1410–1413 (2007)
14. Lugger, M., Yang, B.: Classification of different speaking groups by means of voice quality parameters. In: Proc. of Sprach-Kommunikation (VDE) (2006)
15. Kane, J., Gobl, C.: Identifying regions of non-modal phonation using features of the wavelet transform. In: Proc. of Interspeech (ISCA), pp. 177–180 (2011)
16. Bennett, K.P., Campbell, C.: Support vector machines: hype or hallelujah? ACM SIGKDD Newsletter 2(2), 1–13 (2000)
17. Schwenker, F.: Solving Multi-class Pattern Recognition Problems with Tree-Structured Support Vector Machines. In: Radig, B., Florczyk, S. (eds.) DAGM 2001. LNCS, vol. 2191, pp. 283–290. Springer, Heidelberg (2001)
18. Lin, C.F., Wang, S.D.: Fuzzy support vector machines. IEEE Trans. Neural Net. 13, 464–471 (2002)

Physiological Effects of Delayed System Response Time on Skin Conductance

David Hrabal, Christin Kohrs, André Brechmann, Jun-Wen Tan, Stefanie Rukavina, and Harald C. Traue

Medical Psychology, Ulm University
89069 Ulm, Germany
Leibniz Institute for Neurobiology
39118 Magdeburg, Germany
david.hrabal@uni-ulm.de
http://www.uni-ulm.de,
http://www.ifn-magdeburg.de

Abstract. Research on psychological effects of delayed system response time (SRT) has not lost its topicality, since uncertainty in providing immediate system response remains, even after decades of stunning enhancements in computer science. When delays occur, the user's expectancy about the temporal course of an interaction is not fulfilled which he may interpret as irritating. The current study investigates physiological effects on the skin conductance (SC) and its particular patterns in two experimental scenarios. In the first scenario, unexpected delays of 0.5, 1, and 2 seconds occur while the subject is performing a two-choice auditory categorization task, expecting the system to respond immediately after their input. The second scenario is a wizard-of-oz (woz) scenario in which the user plays the game 'concentration' that is being manipulated in order to induce various emotional states. During the 'negative' sequences delays of 6 seconds are triggered. The patterns of the mean SC curves during delays are analyzed.

Keywords: skin conductance, system repsonse time, emotion recognition, physiological patterns.

1 Introduction

System response time (SRT) research dates back to the late 1960s. But it is still an important issue in computer science research. As today network-based computing gains importance, software engineers have to be aware of network-related delays to be able to improve user performance and satisfaction [1]. Numerous studies concerning the best system response time for a user have been conducted, recommending specific response-time guidelines. These guidelines show that SRT durations of more than a few seconds are accepted when interaction tasks get more complex [2]. However, it is a different situation with very simple, repetitive tasks. These, also called control tasks [1], should behave like physical devices

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 52–62, 2013.

and respond immediately or at least in a few tenths of a second. With more experience in usage, expectations about the SRT begin to establish. While a user is able to adopt to constant delays, variable, unexpected delays often disturb the process of interaction. Especially variations of twice the anticipated SRT might decrease the user's performance and cause frustration [3].

Numerous studies support that an increase in SRT leads to frustration, annoyance and irritation (see [1]). Rating the quality of the system reveals a decrease in perceived quality with increasing SRT. Furthermore, acceptance of such a system decreases. Especially under time pressure users get frustrated, annoyed or even angry by long system response times [3].

The first study investigates whether a small delay of only 500ms is already sufficient to elicit a physiological reaction and whether this response increases with longer delays lasting one or even two seconds. The results are compared to the results of the second scenario, where delays of six seconds are triggered in a stressful situation. Here we also have the possibility to compare the SC patterns during immediate display of a card compared to a delayed display of a card.Here an even greater response in SC is expected.

2 Methods

2.1 Task Description – Experiment 1

32 right-handed subjects (16 female, 16 male) aged 20 to 32 years participated in the first experiment, a simple auditory categorization task with upwards and downwards frequency modulated tones (FM tones). Two subjects were removed due to strong movement artifacts and another seven subjects because their SCR habituated very fast to the presented stimuli. This is not suprising as Venables and Mitchel [] found in their study that nearly 25% of the participtants showed no SCR. Mean age of the remaining twenty-three subjects (10 female, 13 male) was 26 years.

Linearly frequency modulated tones with a duration of 400 ms served as acoustic stimuli. The FM tones differed in direction of frequency modulation (25 upward, 25 downward) and in their center frequency (FC = 1000-3400 Hz in steps of 100 Hz). The 150 different FM tones were presented pseudo-randomly up to two times depending on the participants' overall reaction times. The FM tones were presented with a jittered intertrial interval of two, three or four seconds. The participants' task was to categorize the FM tones according to the direction of modulation. They had to press a button with the right index finger in response to upward modulated FM tones and another button with their right middle finger indicating downward modulated FM tones. During the experiment a white fixation cross on a gray background pointed to the location where the subsequent feedback was displayed for one second; i.e. a green checkmark for correct responses or a red cross for incorrect ones. In the lower right corner of the display a countdown was presented that counted backwards in one-second intervals for one minute. After 30 seconds the digits of the counter turned from white to red, to increase time pressure. The experiment consisted of 20 blocks

of one minute stimulus presentations each. Depending on the the reaction times of the participant up to 15 trials could be solved during one block. In the upper right corner of the display, the remaining number of trials to be solved in the current block was presented. After each block the number of completed trials was presented for five seconds followed by a pause of 25 seconds. In fifteen blocks feedbacks with three different delay-durations (d0.5: 0.5 seconds, d1: 1 second, d2: 2 seconds) served as experimental conditions. All three delays were presented pseudorandomly only once in a block between the 4th and the 12th trial to ensure that most subjects received an equal number of delayed responses. Blocks 1, 4, 8, 14, and 17 contained no delays. Before the experiment, the naïve subjects were not informed about possible delays but were asked to solve as many trials as possible.

During the experiment, subjects were seated in a comfortable chair in a room shaded from daylight. The average temperature was 24°C (±1.5°C) and the average humidity was 49% (±2%). After the instruction and the adjustment of stimulus loudness to a comfortable level, participants were connected to the physiological setup. The silver/silver chlorid electrodes for measuring the skin conductance were placed on the distale phalanx of the forefinger and ring finger of the left hand (SC/GSR Sensor, Nexus-32, Mind Media, The Netherlands) and the peripheral blood volume sensor (BVP-Sensor, Nexus-32, Mind Media, The Netherlands) on the fingertip of the middle finger. The sampling rate of all physiological measurements was set to 512 Hz. The dynamic of the button press of the participants' right index- and middle finger was recorded in steps of two ms using the COVILEX ResponseBox 2.0 (COVILEX GmbH, Germany). After a short rest period of three minutes the participants started the experiment, which lasted about half an our, by pressing a button. The synchronous recording of all physiological and behavioral data was started by a trigger signal while the first button press of the subject served as a reference signal.

2.2 Task Description – Experiment 2

In the second experiment, a wizard-of-oz (woz) experiment, the participants had to solve a memory training task. The concept of woz experiments is widely used for software development and prototyping in the area of Human-Computer-Interaction and interface design [4], [5], [6], [7].

The subject is told that he is interacting with the computer and is unaware of the fact that the experiment is being controlled or manipulated by the so called 'wizard'. In this woz experiment, the subject was told that he will perform a sequence based memory test with an autonomous computer system, which is controlled by voice. In each of the six experimental sequences (es01 - es06) a number of hidden pictures was presendted to the subject. His task was to uncover all matching card pairs. The whole interaction was controled by voice. Each experimental sequence was designed to push the subject into a target emotion.

Fig. 1. A subject performing the mental trainer experiment. The parameters SC, BVP, EMG and EEG are measured. The screen shows the experimental sequence two (es02) with two pairs of cards already discovered.

The experimental sequences and the target emotions are shown in figure 2. A subject and his view onto the screen can be seen in figure 1. The induced emotion depended on the following factors:

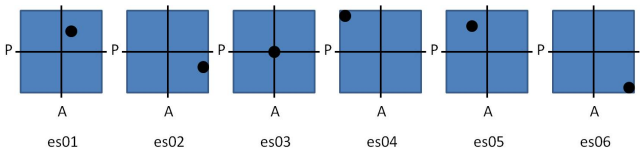


Fig. 2. The target emotions for each of the experimental sequences in the two dimensions of pleasure and arousal

- size of the picture matrix
- alikeness of the pictured motives
- time preassure
- indicated rating of the participant's performance
- positive and negative comments of the system
- incorrect recognition of the subject's commands
- delays in execution of the subject's commands

The subject's screen during experimental sequences four (es04) and six (es06) of the woz experiment can be seen in Figure 3. In es04, the pictures looked very much alike, there was time preassure and the performance bar showed 'under average'. The induced emotion was low pleasure and high arousal (LPHA). In es06, the pictures were different, there was no time preassure, positive comments were

given to the subject and the performance bar showed 'very good'. The induced emotion was high pleasure and low arousal (HPLA). After the experiment, a manipulation check was assessed by a semi-structured interview which included questions about the subject's feelings in all six experimental sequences of the woz experiment and a rating of these sequences using the Self Assessment Manikin (SAM) [8].

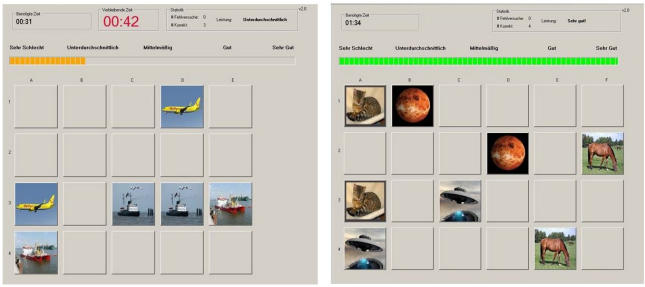


Fig. 3. The subject's screen during experimental sequences 4 (es04) and 6 (es06). The induced emotions were low pleasure and high arousal 'LPHA' in es04 (left) and high pleasure and low arousal 'HPLA' in es06 (right).

Physiological data from 119 subjects (78 female, 41 male) who took part in the wizard-of-oz experiment was analyzed in this study. Mean age was 47 years (SD = 23.27). Altogether 442 cases of delayed display of a card and 442 cases of immediate display of a card were analyzed.

In the woz experiment, the participants had to solve a mental training task familiar to the game 'concentration'. In each of the six experimental sequences (es01-es06) a number of hidden pictures was presented. The subject's task was to uncover all matching card pairs. The whole interaction was controlled via voice. In es04, among other tools, delays were used in order to induce negative emotions.

3 Data Analysis

3.1 Experiment 1

The physiological skin conductance data was imported into MATLAB and down sampled to 16 Hz. The data was analyzed with *Ledalab* ([9]). Deconvolution analysis was performed separating the SC data into continuous signals of tonic and phasic activity [9]. The skin conductance response was calculated by averaging the time range of second two to second four after each button press. The phasic SC response is described by changes in amplitude higher than 0.01μS. All scores were standardized with the formula $y = log(1 + x)$ to account for the positively skewed distributions of SCR amplitudes [10]. The mean SCR was calculated for the following conditions

– immediate feedback
– delayed feedback

and for the three different delays of 0.5, one or two seconds (d1, d2, d3). Data was tested for normal distribution (Kolmogorov-Smirnov-Test; [11]). A general linear model with repeated measurement was computed. The effect of the three different delays on skin conductance was analyzed by computing a polynomial trend test with the factor delay consisting of four steps: immediate feedback (d0), 0.5 seconds (d1), one second (d2), or two seconds (d3). The subsequent post-hoc tests (Boferroni corrected) facilitate the interpretation of trend tests.

3.2 Experiment 2

There are two paradigms in the second experiment:

– The card is displayed approximately one second after the user's request (immediately)
– The card is displayed approximately seven seconds after the user's request (delayed)

An overview over the time course of the experiment and in the paradigms 'no delay' (d = 0) and 'delay' (d = 6) can be seen in Fig. 4. The raw data was imported into MATLAB and for each case a baseline was calculated taking the mean SC value of one second before the user's request for the display of a card. This baseline was then subtracted from the next five seconds, ranging from the user's request for display of a card to five seconds after the request (see figure 4). In the case of immediate display, the card was displayed approximately one second after the request. In the case of a delayed display (in altogether 442 cases), the card wasn't displayed until two seconds after the analyzed range.

After the experiment a SAM (Self Assessment Manikin) rating [8] was surveyed to gain emotional ratings of all six experimental sequences of the woz experiment.

4 Results

4.1 Experiment 1

To compare the effects of the four different durations of a delay (d0, d1, d2, d3) on the skin conductance, a polynomial trend test was computed. Mauchly's test indicates that the assumption of sphericity for the factor delay was violated, $\chi^2(5) = 13.08; p < 0.05$. Therefore, degrees of freedom were corrected using the Greenhouse-Geisser estimates of sphericity ($\epsilon = 0.71$). All four 'delay' versions differed significantly from each other $F(3.66) = 4.12; p < 0.05$. There was a significant linear trend $F(1.22) = 8.30; p < 0.01$ and a significant quadratic trend $F(1.22) = 5.53; p < 0.05$. This indicates a logarithmic increase of skin conductance response as the duration of an unexpected delay increases. The post

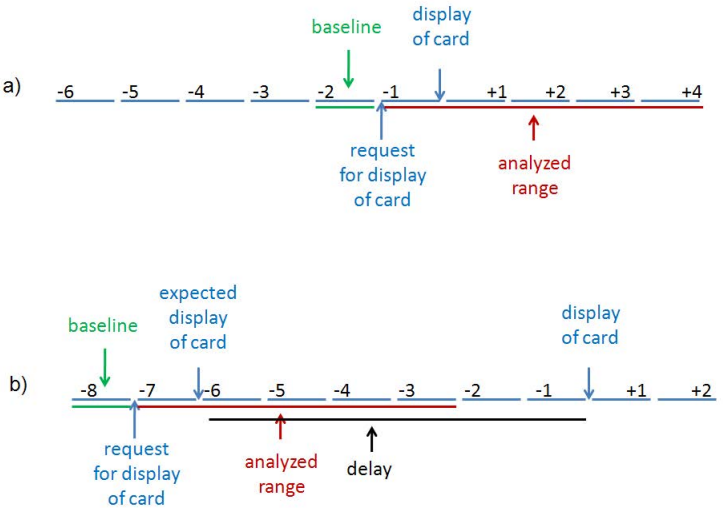


Fig. 4. The time course in the two cases of immediate (a) and delayed (b) display. In (a), the card is displayed aproximately one second after the subject's request. In (b), there is a delay of six seconds. The SC range plotted in Fig. 8 ranges from the user's request for the display of a card to five seconds after the request.

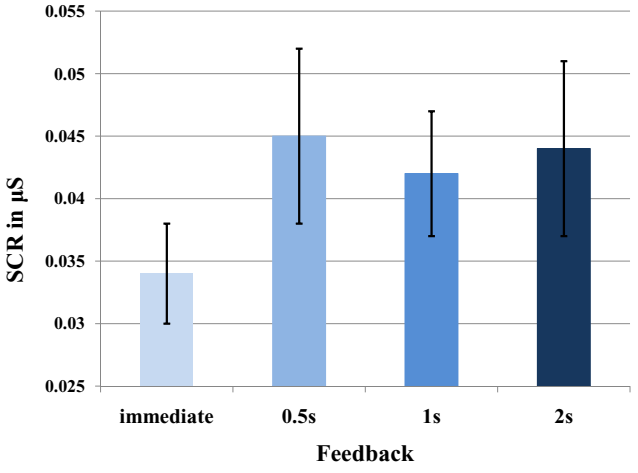


Fig. 5. Mean and standard error of the SCR for the conditions d0, d1, d2 and d3. The SC response to immediate feedback is significantly lower than the response to delays of 0.5, 1 or 2 seconds.

hoc tests revealed a significant difference between immediate feedback (d0) and d2 and d3, $p < 0.05$ (Bonferroni corrected). The difference between d0 and the shortest delay (d1) was not significant. Furthermore, there was no significant difference between the three delays d1, d2 and d3, $p = 1$ (see Fig. 5).

The mean increase of 0.009 μSiemens in the button press experiment comparing an immediate feedback with the delayed feedbacks corresponds to a gain of 26.47% in the SCR.

4.2 Experiment 2

The course of the SC curves shown in Fig. 8 illustrates the mean of all 442 SC curves from all 119 subjects during a delay (red) and 442 SC curves from all 119 subjects during immediate display of a card (green) preceding the delayed displays.

As can be seen in Fig. 8 the divergence of the SC cureves begins approximately 2.3 seconds after the display of a card / expected display of the card. It reaches its maximum (0.0125 μSiemens) at the end of the analyzed range of 5 seconds.

The mean SAM ratings for the dimension 'pleasure' were 5.5 for es04 and 4.0 for es05. These were the experimental sequences where the delay was applied. For the sequences es01, es02, es03 and es06 the mean SAM ratings were 7.6, 7.4, 7.2 and 7.8.

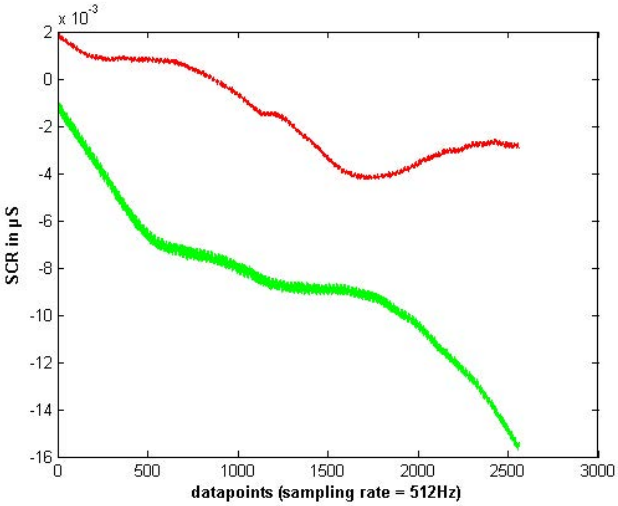


Fig. 6. Plot of the mean SC reaction of 119 subjects to alltogether 442 delayed card openings (red) and 442 immediate card openings (green) during the memory training experiment. The baseline was substracted from each point.

The mean difference in SC behaviour between the trials immediate and the delayed trials over the range of five seconds is 0.007 μSiemens. Assuming a significance level of 0.05 the two conditions differ significantly (paired t-test, p-value <0.00001).

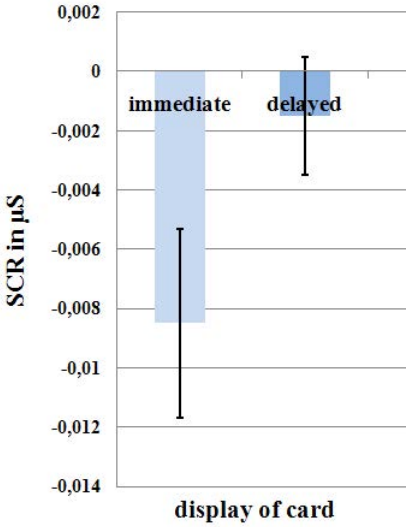


Fig. 7. Mean and standard error of the SCR for the conditions of immediate card display and delayed card display (t = 6s). The SC response to immediate display is significantly lower than the response to delayed display.

5 Discussion

The physiological data of experiment 1 reveals an increase in skin conductance during unexpected delays compared to immediate feedback presentation. Even delays of only 500ms are sufficient to trigger this physiological change, but longer delays elicited a greater response. The increase of skin conductance is in line with findings of [12]. They found a greater number of spontaneous SCRs and an elicited skin conductance level during blocks of longer SRTs (8s) compared to blocks of shorter SRTs (2s). The results of the current study reveal that even very short delays in SR which are still common in HCI, unsettle the participant about the further temporal course of the interaction. When unexpected delayed system responses occur the user is faced with two possible questions: 'was may action registered?' and 'do I have to repeat it again?' [13]. This uncertainty about the further temporal course of the interaction increases cognitive and/or emotional demands and may elicite an orienting response of the organism with its typical physiological changes like an increased SCR [14].

Experiment 2 has a continuative design with embedded natural dialog. Here we could observe the effects of delays in a stressful realistic scenario. The observed difference in the averaged SCL is even greater than in experiment 1. This can be explained by the setup of the experiment. The lack of display of a requested card is a negative event for the subject particularly in already very stressful situations like the experimental sequence 4 of this experiment.

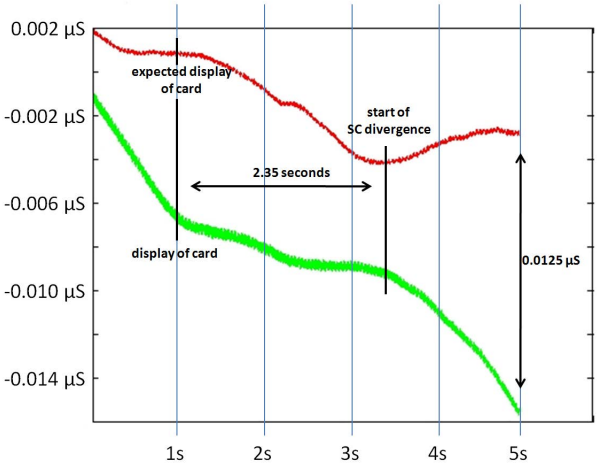


Fig. 8. Analysis of the plot of the mean SC reaction of 119 subjects to alltogether 442 delayed card openings (red) and 442 immediate card openings (green) during the memory training experiment. The divecrgence of the SC curve begins around 2.35 seconds after the diplay of a card / expected display of a card.

The pattern of the SC curve changes 2.35 seconds after the display of a card / expected display of a card. With the latency of a skin conductance reaction being between 1.0 and 3.0 seconds ([15]), the introduced delay perfectly explains the change of the SC pattern after 2.35 seconds.

Since nearly all subjects reported negative emotions elicited by the delays in the SAM ratings, the rise of skin conductance level can, in this case, be interpreted as negative arousal and should be avoided in human computer interaction.

Acknowledgments. This research was supported by grants from the Transregional Collaborative Research Centre SFB/TRR 62 'A Companion-Technology for Cognitive Technical Systems' founded by the German Research Foundation (DFG).

References

1. Dabrowski, J., Munson, E.V.: 40 years of searching for the best computer system response time. Interacting with Computers 23, 555–564 (2011)
2. Shneiderman, B., Plaisant, C.: Quality of services. In: Designing the User Interface - Strategies for Effective Human-Computer Interaction, 4th edn., pp. 453–475. Pearson Addison Wesley, Boston (2005)
3. Shneiderman, B.: Response time and display rate in human performance with computers. ACM Comput. Surv. 16, 265–285 (1984)

4. Hoysniemi, J., Hamalainen, P., Turkki, L.: Wizard of oz prototyping of computer vision based action games for children. In: IDC 2004: Proceedings of the 2004 Conference on Interaction Design and Children, pp. 27–34. ACM, New York (2004)
5. Molin, L.: Wizard-of-oz prototyping for co-operative interaction design of graphical user interfaces. In: NordiCHI 2004: Proceedings of the Third Nordic Conference on Human-Computer Interaction, pp. 425–428. ACM, New York (2004)
6. Bernsen, N.O., Dybkjaer, H., Dybkjaer, L.: Wizard of oz prototyping: How and when? In: CCI Working Papers in Cognitive Science and HCI (1994)
7. Akers, D.: Wizard of oz for participatory design: inventing a gestural interface for 3d selection of neural pathway estimates. In: CHI 2006: CHI 2006 Extended Abstracts on Human Factors in Computing Systems, pp. 454–459. ACM, New York (2006)
8. Bradley, M., Lang, P.: Measuring emotion: the self-assessment minikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry 25(1), 49–59 (1994)
9. Benedek, M., Kaernbach, C.: A continuous measure of phasic electrodermal activity. Journal of Neuroscience Methods 190(1), 80–91 (2010)
10. Venables, P.H., Christie, M.J.: Electrodermal activity. In: Martin, I., Venables, P.H. (eds.) Techniques in Psychophysiology (1980)
11. Massey, F.J.: The kolmogorov-smirnov test for goodness of fit. Journal of the American Statistical Association 46, 68–78 (1951)
12. Kuhmann, W., Boucsein, W., Schaefer, F., Alexander, J.: Experimental investigation of psychophysiological stress-reactions induced by different system response times in human-computer interaction. Ergonomics 30(6), 933–943 (1987)
13. Pérez-Quiñones, M.A., Sibert, J.L.: A collaborative model of feedback in human-computer interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground, pp. 316–323. ACM, New York (1996)
14. Ben-Shakhar, G.: The roles of stimulus novelty and significance in determining the electrodermal orienting response: Interactive versus additive approaches. Psychophysiology 31, 402–411 (1994)
15. Cacioppo, J.T., Tassinary, L.G., Berntson, G.: The Electrodermal System. In: Handbook of Psychophysiology, 168, Cam (2007)

A Non-invasive Multi-sensor Capturing System for Human Physiological and Behavioral Responses Analysis

Senya Polikovsky, Maria Alejandra Quiros-Ramirez, Takehisa Onisawa, Yoshinari Kameda, and Yuichi Ohta

Graduate School of System and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8573 Japan
senya@image.iit.tsukuba.ac.jp

Abstract. We present a new noninvasive multi-sensor capturing system for recording video, sound and motion data. The characteristic of the system is its 1msec. order accuracy hardware level synchronization among all the sensors as well as automatic extraction of variety of ground truth from the data. The proposed system enables the analysis of the correlation between variety of psychophysiological model (modalities), such as facial expression, body temperature changes, gaze analysis etc... . Following benchmarks driven framework principles, the data captured by our system is used to establish benchmarks for evaluation of the algorithms involved in the automatic emotions recognition process.

Keywords: sensor-fusion, synchronization, benchmarks.

1 Introduction

Automatic recognition of emotions has been actively studied in the last decade [7]. Although strong benchmark environment is necessary for the development of this field it is usually neglected. In this work we present a new noninvasive multi-sensors capturing system for collecting video, sound and motion data that allows the creation of benchmarks. The recorded multi-sensor data enables the analysis of correlation between diverse psychophysiological models, such as facial expression, body temperature changes, gaze analysis, and voice.

Psychophysiological models are usually studied independently and fusion between new sensing technologies is barely utilized. There are four reasons for limited use of sensor fusion: first, there is an absence of a single off-the-shelf system that integrates a variety of modern sensors and simplifies their manipulation. Second, having several recording devices brings up the challenge of their synchronization. Synchronization is a key point in order to analyze emotional changes in time and relation between different clues in representing emotions. Third, the overflow of the recorded video and other data makes its management and analysis difficult. Finally, in order for the recorded data to be used as a benchmark for tracking and classification algorithm development, a variety of

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 63–70, 2013.

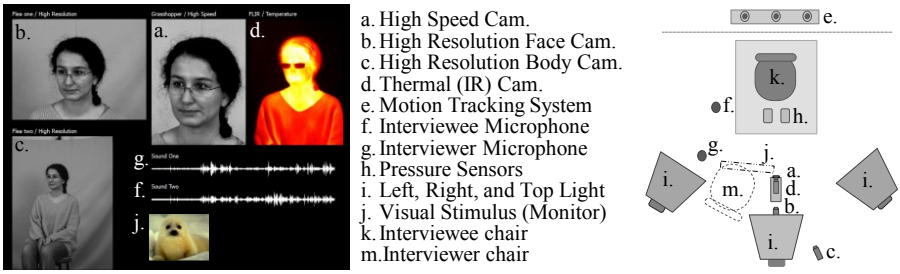


Fig. 1. Left: Visualization of the captured data during Computer-to-Human interview by our system. Right: Diagram of the sensors location during data collection.

ground truth information should be annotated which represents a challenging and time consuming task.

The capturing system we present contains high speed, high resolution and thermal cameras, eyes and 3D marker tracking devices, microphones and pressure sensors. The system provides a simple control over sensors and ensures 1msec order accuracy hardware level synchronization among all the sensors. The design of the system allows relatively simple extensibility of additional sensors. In addition, we introduce sensors for speeding-up the annotation process by segmenting points of interest in recorded data as well as extracting ground truth information such as position of the body parts and gaze direction. These characteristics are required for the creation of benchmarks. The system was used to record a cross-cultural database containing 36 subjects, presented at [11]. Figure 1, left side shows an example of signals captured by the system and right side is a diagram of the sensors location during the recording.

Based on our benchmark driven framework [2] used for development of emotion sensing systems, the presented system corresponds to the "Data Capturing" step of the framework. The data captured by the system is used to create the benchmarks of the remaining steps of the framework. This topic will be explained in detail in the section 3.

There is few research focused on the design of capturing systems. The research presented in [5] is an example of some guidelines for designing a capturing system with special emphasis on synchronization between video, sound and eye tracking sensors. In this paper we present a system design that implements a broader range of sensors utilizing a different design approach.

Due to the variety of terms in the field, we chose *psychophysiological model* to refer to similar terms such as behavioral clue or modality.

2 Multi-sensors Capturing System

This section introduces the design of our multi-modal capturing system. The design of a capturing system for emotional sensing purpose consists of the following steps: 1) Definition of the scenarios in which the system will be used.

2) Choice of the psychophysiological models that will be analysed. 3) Selection of system sensors based on measurement accuracy, synchronization capabilities, sensor hardware and software interfaces, as well as the price. 4) Choice of hardware configuration for system computers, operation system, development environment and communication protocols in the system. During the design process the steps go through a number of iterations until the final configurations are established. From our experience due to the large number of factors and limited guidelines on capturing system design, the process is more empirical than methodological. Therefore an introduction of a variety of capturing system designs is necessary for advancing the field since it will save the development time that it requires to build the system from scratch.

Next we describe the scenarios and the selected psychophysiology models, then we define the requirements of the capturing system and present the system design. Finally, we introduce the system sensors and synchronization scheme.

2.1 Scenario and Psychophysiology Models

In this stage we are focusing our interest on human-to-human and human-to-computer indoor sitting interview scenario. This scenario allows to control environmental factors such as lighting conditions, room temperature and space background. The use of the chair limits the movement of the interviewee and dictates sensors location.

As for the psychophysiology models, based on collaborations with a police negotiation unit as well as a psychologist [3] from Arizona University, we have identified the five most promising psychophysiology models for behavior analysis from a technological and psychological point of view. These models are used to differentiate between normal and abnormal behavior, detection of deception and stress.

1. Micro-expressions - Ekman at al. showed that facial expressions are the most important behavioral source for lie indication and danger demeanor detection [1]. Micro-expressions appear with low muscular intensity, which makes it impossible to analyse using standard speed cameras. Thus, a high speed camera is required [4].
2. Facial Feature Area Temperate - Pavlidis at al [6]. demonstrated the correlation of increased blood perfusion in the orbital muscles and stress levels for human beings. It has also been suggested that this periorbital perfusion can be quantified through processing thermal video captured by thermal (infrared) camera.
3. Eyes analysis - Pupil dilation as well as gaze direction [10] can also be used for stress, interest and drug use detection. The newest eye tracking systems provide high accuracy analysis.
4. Body Language - Analysis of head, shoulders and hands movement can be used for deception detection. Body language has been used for decades by psychologists for human behavior analysis [3].
5. Voice Stress Analysis - Used to recognize stress responses that are present in human voice, when a person suffers psychological stress [8].

By combining these five approaches that rely on different sources of information, we increase trustfulness and robustness of the final analysis. In addition, the accurate synchronization between the measurements related to each one of the models provides the ability to analyse the timing correlation between them. The synchronization of 1msec order was selected as a trade-off between sufficient accuracy (that allows to combine EMG signals in the future) and the cost and complexity of the system.

2.2 Capturing System Requirements and Design

The system was designed based on the following list of requirements: 1) all sensors are controlled using a simple interface from a single computer, 2) all sensors have the ability to be synchronized with 1msec order accuracy, 3) simple integration of new sensors is allowed, 4) video data is captured with no compression, allowing analysis of the influence of compression in the future, 6) detection of missing frames in video sequence is supported, 7) automatic extraction of stimulus timing to speed-up the segmentation of the recorded signals, 8) automatic extraction of ground truth by off-the-shelf devices, 9)system is capable of recording large amount of information into the HDDs, 10) system is transportable.

Figure 2 introduces the capturing system design, due to the limitation of the space the overview of the design is the caption of the figure. For more details on the design we encourage the reader to contact the authors.

2.3 Sensors

The sensors of the system can be separated in two groups: the first group contains the sensors that will be used in real-time implementation; the second group contains support sensors that are aimed to extract reliable ground truth measurements. Some of the sensors belong to both groups (see Table 1). The sensors from the first group are: high-speed camera for analysis of facial micro-expressions, infrared camera measuring temperature changes, high-resolution camera for body analysis and speaker interaction, microphones capturing voice, pressure sensors for capturing body weight changes and rapid legs motion. The second group contains motion capturing system for automatically detect precise head location and orientation as well as shoulder level. The MCS markers are attached on the backside in such way that they cannot be seen by front cameras. Eye tracking system for automatic extraction of gaze and pupil dilatation. Photosensor for capturing precise timing of the visual stimulus presented on the screen.

A photosensor with response similar to human eyes is attached to the computer screen to detect the exact timing of visual stimulus presented on the screen during human computer experiments. Having the exact timing of the stimulus allows the automatic segmentation of the recorded data in order to extract the important sections and reduce the amount of final data.

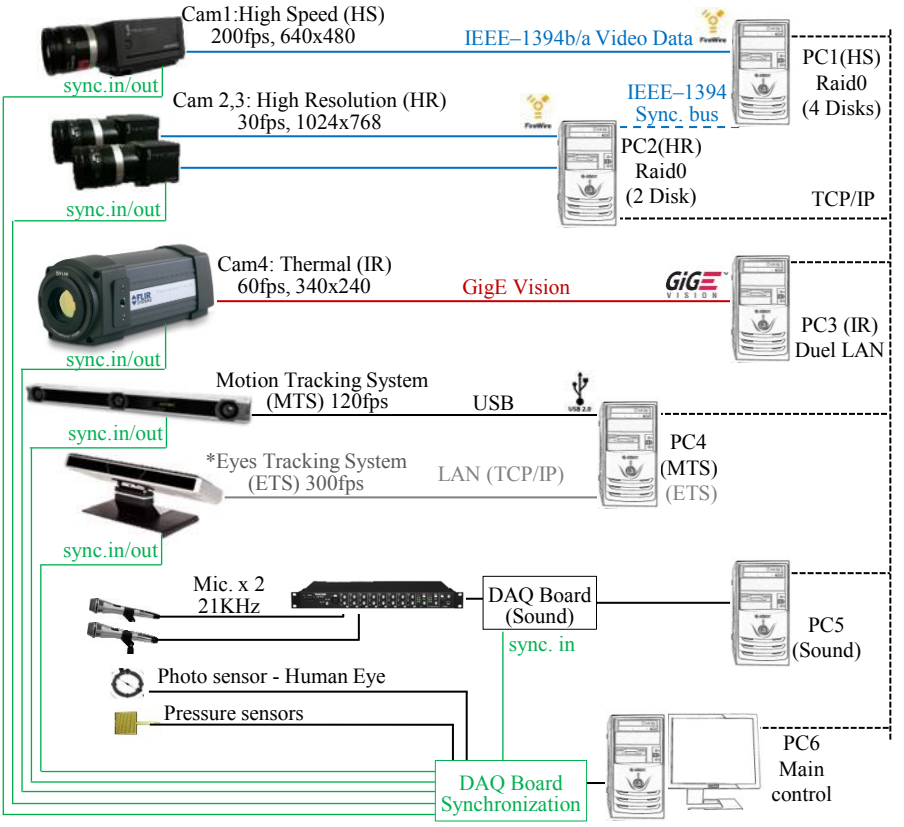


Fig. 2. Capturing system design. The input sensors are: high speed, thermal, and two high resolution cameras, motion and eyes tracking systems, two microphones, pressure sensors, and a photosensors with a response similar to the one of human eyes. The corresponding cameras' data transfer interfaces can be seen in the figure. Due to 1) incapability of number of cameras' drivers to be install on the same computer, 2) reuse of already exiting equipment and 3) a large capacity of the recorded data, separate computer is dedicated for each on of the sensor.

The system consists of six computers connected to the same network, PC1 and PC2 contains RAID0 hard disk setup for high speed stream data recording. Special attention should be given to configuration of the RAID hardware to meet the required writing speed. For HS and HR cameras, the use of HPwv8400 workstation with onboard RAID card provided the sufficient writing speed. The rest of the computers are standard configuration PC. The main computer controls the simultaneous recording process of all the sensors through MailSlot interprocess communication. Computers that receive data from 1394 protocol have an additional synchronization 1394 bus (dash blue line). PointGrey provides synchronization software on top of 1394 bus, however it supports only on WinXP OS, therefore WinXP and Win7 are used as OS in the system. All the sensors are wired with synchronization in/out cable (green line) connected to data acquisition board (DAQ), additional synchronization signals are supplied by a sound recording system. Photo and pressure sensors and sampled using the same DAQ, this way all the digital and analog signals are recorded in the same time line.

* At this time Eyes Tracking System was not used during the data collection.

Table 1. System sensors parameters

Sensor	fps	Data type	Sync	role
Hi-speed (Grasshopper, PTG)	200	640x480, RGB	GPIO	on-line
High-resolution (Flea2, PTG)	30	1024x768, RGB	GPIO	on-line/support
Thermal (A325, FLIR)	60	320x240, 16 bits	GPIO*	on-line/support
Motion Tracking System (Trio)	120	3D points location	GPIO**	support
Eyes Tracking System (TobiiTX300)	300	Gaze direction	GPIO	on-line/support
Microphones x2	21(KHz)	1D	Acq.Card	on-line/support
Photodiode	2(KHz)	1D	Acq.Card	on-line/support
Pressure sensors X4	2(KHz)	1D	Acq.Card	on-line/support

PTG - Pointgrey, * - No Shutter out, ** - Start / stop control,
Tobii - sync. option coming soon

2.4 Synchronization

In this section we describe our synchronization strategy and alignment of the recorded data. Current off-the-shelf capturing products are not capable to ensure high accuracy synchronization between cameras based on different interfaces such as FireWire (IEEE 1394) and GigE Vision and Camera Link. One possible solution is to use an external trigger; however high-speed middle range price cameras lack accuracy in external trigger mode or not support it as with thermal cameras. The high-end price of such cameras can reach order of 100k$. In our strategy the main computer, that controls the operation of all the sensors, starts the recording of all sensors asynchronous. Next, after insuring that all the sensors

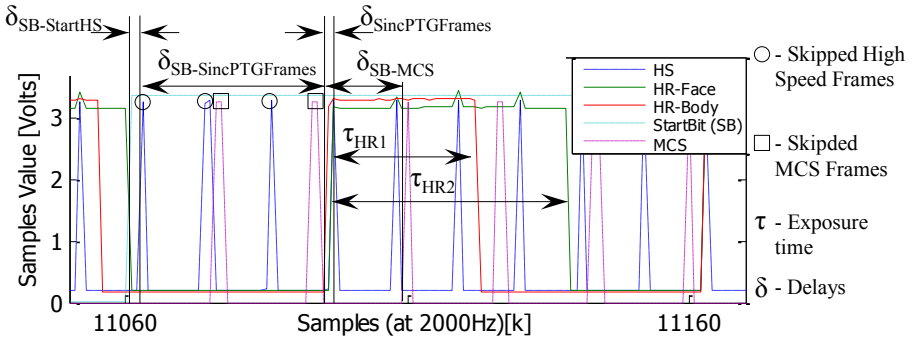


Fig. 3. Cameras strobe signals captured by DAQ that are used to extract the time delay between the sensors. The delays are calculated in correspond to "start bit" (SB) (dash green). $\delta_{SB-FirstHS}$ delay between SB and start frame of high speed camera. $\delta_{SB-SincPTGFrames}$ - SB and the first time all the frames of PTG cameras are sync. $\delta_{SincPTGFrames}$ - PTG cameras sync. error (500msec.). δ_{SB-MCS} - SB and motion capturing system frame. ◯ and □ - indicates the HS and MCS frames that should be skipped during the alinement process. τ represents the cameras exposure time.

are recording, data acquisition board (DAQ) sends the "start" bit to general purpose I/O (GPIO) port of all the sensors. As the state of GPIO port values are embedded into the frames, this way we mark the start frame to use in alignment process. (The stop of the recording done in the same way, first sending the "stop" bit and then ending the recording asynchronous)

In order to implement this approach all the sensors should be capable to embed the fames ID or camera internal timestamp, and GPIO port state into the recorded frames. The use of frames ID allow automatic detection of missed (unrecorded) frames. Besides, if the camera provides an output strobe pulse corresponding to shutter manipulation, it allows to calculate the time delays between the start bit and the frames, to increase the synchronization accuracy (see figure 3). To synchronise between PointGrey cameras we use the build-in sync. option on top of FireWire, however due to differences in cameras fps the synchronization accuracy increases form 0.125msec. to about 0.5msec. In [9] we provided an evaluation of the proposed synchronization strategy accuracy between high speed and thermal cameras.

3 Capturing System for Benchmark Driven Framework

The purpose of a benchmark driven framework in the field of emotion recognition is to have an effective collaboration platform between technological and psychological researches as well as intensive benchmark capabilities to test the performance of the entire system as well as individual algorithms. The core of this framework is the carefully prepared benchmarks that correspond to the following four steps: "Data Capturing", "Feature Tracking", "Feature Analysis" and "Classification".

Our capturing systems implements the first step of this framework. We assure that the recorded signals will be automatically temporally aligned, segmented and the available ground truth will be extracted. Synchronization of the signals plays a key role for temporal alignment. It is necessary to remark that the lack of any of the characteristics of our capturing system would complicate the use of the benchmark driven framework.

Using the framework to support the development of emotion recognition system allows the evaluation of different algorithms in each one of the steps that follow the data capturing, as well as measuring the significance of each of the steps. These are two characteristics that have been left aside at the time of designing and building emotion recognition systems nowadays.

4 Conclusion

In this paper we introduce a multi-sensor non-invasive capturing system that will provide new, previously not available, sources of information for physiological and behavioral researchers. The important features of the system are: First, simple manipulation of multiple sensors. Second, high accuracy synchronization between the sensors that allows to analyze psychophysiology modal correlation

and to develop more robust and reliable models for human emotion detection. Third, the system automatically segments points of interest in recorded data. Finally, it provides hardware support for faster extraction of ground truth from the recorded videos.

Current studies in emotion recognition are based on applying tracking algorithms that fit the available data and the use of classification algorithms on extracted information. This empirical approach makes comparisons and analysis of the output results difficult. The benchmark framework approach supports comparison of several tracking and classification algorithms in order to understand their individual and coupled effect over the final emotion classification.

The capturing system we present in this study has been designed in order to support the creation of benchmarks and initiate a process of comparisons towards a methodological way of creating emotion recognition systems in order to leave behind the usual empirical solutions employed up to now.

References

1. Eckman, P.: Telling lies, 2nd edn. Norton (2009)
2. Polikovsky, S., Quiros-Ramirez, M.A., Kameda, Y., Burgoon, J., Ohta, Y.: Benchmark Driven Framework for Development of Emotion Sensing Support Systems. In: Workshop on Innovation in Border Control (2012)
3. Jensen, M.L., Meservy, T.O., Burgoon, J.K., Nunamaker, J.F.: Video-based deception detection. In: Intelligence and Security Informatics, pp. 425–441 (2008)
4. Polikovsky, S., Kameda, Y., Ohta, Y.: Evaluation of synchronization accuracy between high speed cameras in infrared and visible spectrums. In: 3rd International Conference on Imaging for Crime Detection and Prevention (2009)
5. Lichtenauer, J., Shen, J., Valstar, M.F., Pantic, M.: Cost-effective solution to synchronised audio-visual data capture using multiple sensors. Image and Vision Computing 29, 666–680 (2011)
6. Pavlidis, I.T., Frank, M.G., Ekman, P.: Imaging facial physiology for the detection of deceit. International Journal of Computer Vision 71(2), 197–214 (2001)
7. Calvo, R., Mello, S.D.: Affect detection: an interdisciplinary review of models, methods and their applications. IEEE Transactions on Affective Computing 1, 18–37 (2010)
8. Ruiz, R., Legros, C., Guell, A.: Voice analysis to predict the psychological or physical state of a speaker. Aviation Space and Environmental Medicine 61(3), 266–271 (1990)
9. Polikovsky, S.: Evaluation of synchronization accuracy between high speed cameras in infrared and visible spectrums. In: Proceedings of IAPR Conference on Machine Vision Applications, pp. 51–54 (2010)
10. Zhai, J.: Stress detection in computer users through non-invasive monitoring of physiological signals. Biomed. Sci. Instrum. 45, 495–500 (2006)
11. Quiros-Ramirez, M.A., Polikovsky, S., Kameda, Y., Onisawa, T.: Towards Developing Robust Multimodal Databases for Emotion Analysis. In: Proc. of 6th SCIS-ISIS (2012)

3D Motion Estimation of Human Body from Video with Dynamic Camera Work

Matsumoto Ayumi[1], Wu Xiaojun[2], Kawamura Harumi[1], and Kojima Akira[1]

[1] Nippon Telegraph and Telephone Corporation
{matsumoto.ayumi,kawamura.harumi,kojima.akira}@lab.ntt.co.jp
[2] NTT Communications Corporation
x.wu@ntt.com

Abstract. Occlusion or camera setting produces a high degree of ambiguity when estimating human body motion from monocular video sequences. Good human motion models are an important means of addressing this problem. In this work, we propose a hierarchical motion model and a motion estimation for it to estimate human motion without camera calibration and with free camera operation. The model is able to generate particles in multi-spaces and thus is able to estimate both camera view and human motion at one time. We showed the possibility of achieving 3D motion estimation for simple movements such as "walking" without camera calibration and with dynamic camera operation.

Keywords: 3D pose estimation, vision based motion estimation, calibration free, monocular camera.

1 Introduction

Human motion estimation is an important topic in the computer vision field for understanding of human behavior. Many applications, including supervision, human interaction, animation generation and sports science are expected[1,2]. Human motion estimation incorporates pose estimation and motion cognition. The former means computing 3D positions of human joints and the latter means understanding the category of motion. Depth camera [3] and multi-camera[4] systems have been used for pose estimation, but installation (camera calibration and location) and cost limitations restrict their actual use. In addition, these systems do not understand the category of motion.

Our goal is to achieve 3D pose estimation and motion cognition from monocular video sequences with easy installation (i.e., without camera calibration and location limitations). Occlusion or camera setting produces a high degree of ambiguity when estimating human body motion from monocular video sequences. Good human motion models are an important means of addressing this problem. In this work, we propose a hierarchical motion model and a motion estimation method for it to estimate human motion with free camera setting and operation.

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 71–78, 2013.

2 Related Works

Time series data as human motion is modeled by a dynamical system. A state space model is a basic model to explain a standard dynamical system in order to model time series data. We can generally consider a latent variable mapping with first-order Markov dynamics(Fig.1(a)), where f, g are usually defined as follows.

$$\mathbf{x}(n) = f(\mathbf{x}(n-1)) + \eta_x(n) \tag{1}$$

$$\mathbf{y}(n) = g(\mathbf{x}(n)) + \eta_y(n) \tag{2}$$

Here, $\mathbf{x}(n)$ denotes the hidden state of the system at time n, $\mathbf{y}(n)$ denotes the output of the system at time n and $\eta_x(n), \eta_y(n)$ are system noises. Eq.1 therefore shows the mapping the hidden state variable from time $n-1$ to n, and Eq.2 shows the mapping from the state variable to the output variable. This formation is able to predict the present hidden system state from the present output and previous state.

Human motion data consists of human joint angles with about 50 degrees of freedom. An important study point is to construct an essential and general model from high-dimensional motion data. Gaussian Process Dynamical Models (GPDMs) [5] are useful for achieving nonlinear dimensional reduction for time series analysis, comprising a low dimensional latent space with associated dynamics and a map from the latent space to an observation space. A GPDM for human motion [6] has been proposed for estimating human motion from monocular video scenes with occlusion [7].

3 Hierarchical Motion Model

3.1 Key Idea

When a standard dynamical model(Eq.1,2) is applied to human motion modeling for video-based human motion estimation, $\mathbf{y}(n)$ is 3D human motion data such as motion capture data. Our goal is to get 3D human motion data from the observable 2D video data. The 2D mapping parameter is needed to do so, but it is generally unknown. The parameter to 2D mapping is needed in the case, but it is generally unknown. We therefore propose an expanded model, which is shown in Fig.1. Let $\mathbf{X}_{all}, \mathbf{Y}_{all}$ be the same as $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N], Y = [\mathbf{y}_1, ..., \mathbf{y}_N]$ in Fig.1. We adopted a new system of $\mathbf{Y}_{sub}$, which is defined as observed data in same dimension. Our key idea is to define each relation of multiple $\mathbf{X}_{sub}$ and $\mathbf{X}_{all}$ so that we can construct a hierarchical model that can compute the system output and the observed data in the same dimension without the 2D mapping parameter. We can get 3D human motion data $\mathbf{Y}_{all}$ (i.e., the system all-output data) by computing the observable video data I and the system sub-output data $\mathbf{Y}_{sub}$(Fig.1(b)). Specifically, the relation can be denoted in Eq.3, as:

$$\mathbf{x}_{all}(n) = r(\mathbf{x}_{sub}(n)) \tag{3}$$

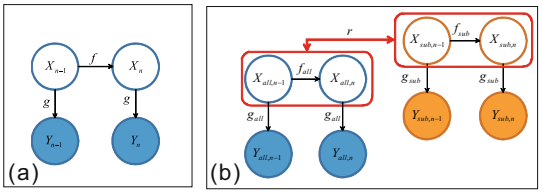


Fig. 1. (a) Graphical model of a standard dynamical system to model time series data with first-order Markov dynamics into the latent space(Eq.1,2). (b) Graphical model of a hierarchical model that can compute the system output and the observed data in the same dimension(Eq.3-5).

3.2 Learning Motion Model with GPDM

GPDM is a latent variable model. It comprises a generative mapping from latent space to data space, and a dynamical model in latent space(Fig.1). It is defined by a set of $q(< D)$-dimensional representations $\mathbf{X} = [\mathbf{x}_1 ... \mathbf{x}_N]^T$ of the D-dimensional data $\mathbf{Y} = [\mathbf{y}_1 ... \mathbf{y}_N]^T$, where N is the number of data sample. We apply GPDM to modeling 3D human motion capture data $\mathbf{Y}_{all}$ and its partial data $\mathbf{Y}_{sub}$, in particular 2D mapping of $\mathbf{Y}_{all}$. That is, we get the whole motion model $M_{all} = [\mathbf{Y}_{all}, \mathbf{X}_{all}, g_{all}, f_{all}]$ and the sub motion model $M_{sub} = [\mathbf{Y}_{sub}, \mathbf{X}_{sub}, g_{sub}, f_{sub}]$. Here, $\mathbf{X}_{all}$ and $\mathbf{X}_{sub}$ denote latent variables, g_{all} and g_{sub} mapping functions from latent space to data space, and f_{all} and f_{sub} dynamical models in latent space.

3.3 Relating Motion Models by Gaussian Process Regression

We want to estimate an unknown latent variable $\mathbf{x}_{all}$ under the condition that we have a known latent variable $\mathbf{x}_{sub}$. This problem can be formulated with Eq.3. A Gaussian Process[8] is applied to a non-linear stochastic process. Essentially, human motions are non-linear in nature. In addition, mapping from partial information to whole information has ambiguities. Therefore, we use Gaussian Process Regression(GP-R) to define regression $\mathbf{X}_{all}$ from $\mathbf{X}_{sub}$.

When learning data are given as $\{\mathbf{X}_{all}, \mathbf{X}_{sub}\}$, GP-R is formulated with Eq.4.

$$p(\mathbf{X}_{all}|\mathbf{X}_{sub} = \prod_{k=1}^{q} N(\mathbf{X}_{all}(:,k)|0, \mathbf{C}_N) \tag{4}$$

$$\begin{aligned}(\mathbf{C}_N)_{i,j} &= C(\mathbf{x}_{sub}(i), \mathbf{x}_{sub}(j)) \\ &= k(\mathbf{x}_{sub}(i), \mathbf{x}_{sub}(j)) + \gamma^{-1}\delta(i,j)\end{aligned} \tag{5}$$

GP-R is defined by the covariance function $\mathbf{C}_N$ or the kernel function $k(\mathbf{x}_{sub}(i), \mathbf{x}_{sub}(j))$ and hyper parameter γ.

Thus, we can get the mapping function r in Eq.3 by maximizing the log posterior of Eq.4.

3.4 Viewing Variation

The data representation of 3D human motion is not a world coordinate, which is a set reference point in real space, but a relative coordinate, which is a set reference point in the human waist position, namely a root the has broad utility. However, in the actual video, we put a camera into real space. Therefore, the relative positions of the camera and the human root change at all times. For this reason, we need to consider that this relative view changes in 3D human motion estimation with actual videos. In this work, we apply a hierarchical motion model to 3D human motion estimation involving relative view changes. Actually, since we build a hierarchical motion model for every view in the learning phase, we consider relative view changes by handling views as state variables in the estimation phase.

Fig.2 is an overview of every-view learning via a view sphere. We assume a view sphere that is a set reference point on the human waist position, then sample the sphere surface at regular intervals. We set virtual cameras at each sampling position, then map 3D human joint positions as 2D screen positions each time. These 2D screen positions are treated as partial motion data in the above hierarchical motion model. We learn the hierarchical motion model for each view $M_{V_i} = [V_i, \mathbf{Y}_{sub}, \mathbf{X}_{sub}, g_{sub}, f_{sub}]$, where the ith sampling position is defined as view $V_i = [\theta_i, \phi_i]$, θ is the angle of orientation, ϕ is the angle of elevation. We need one hierarchical motion model per one motion category such as 'Walk', 'Run' and 'Skip'.

4 Camera Angle Estimation and Motion Tracking Method

4.1 Formulation of State Estimation

The state at frame n is defined as,

$$\Phi(n) = [V(n), \mathbf{y}_{sub}(n), \mathbf{x}_{sub}(n), S(n)] \tag{6}$$

where $V(n)$ denotes the view parameter, $\mathbf{y}_{sub}(n)$ the 2D joint position, $\mathbf{x}_{sub}(n)$ the latent variable, and $S(n)$ the scale parameter in frame n. We estimate a state sequence $\Phi(1:N) \equiv (\Phi(1), ...\Phi(N))$ when given an image sequence $\mathbf{I}(1:t) \equiv (\mathbf{I}(1), ..., \mathbf{I}(n))$ and the learned motion model M_{V_i} in (7),

$$\begin{aligned} &p(\Phi(n)|\Phi(n-1)) \\ &\quad = p(V(n), \mathbf{x}_{sub}(n-1)|V(n-1)) \times p(\mathbf{x}_{sub}(n)|V(n)) \\ &\quad \times p(\mathbf{y}_{sub}(n)|\mathbf{x}_{sub}(n)) \times p(S(n)|S(n-1)) \end{aligned} \tag{7}$$

then get a whole latent variable $\mathbf{x}^*{}_{all}$ and 3D human motion data $\mathbf{y}^*{}_{all}$ with an estimated sub latent variable $\mathbf{x}^*_{sub}$, a learned motion model M_{all}, and the mapping function r in (8,9) .

$$\mathbf{C}^* = (C(\mathbf{x}^*_{sub}, \mathbf{x}_{sub}(1))...C(\mathbf{x}^*_{sub}, \mathbf{x}_{sub}(n))) \tag{8}$$

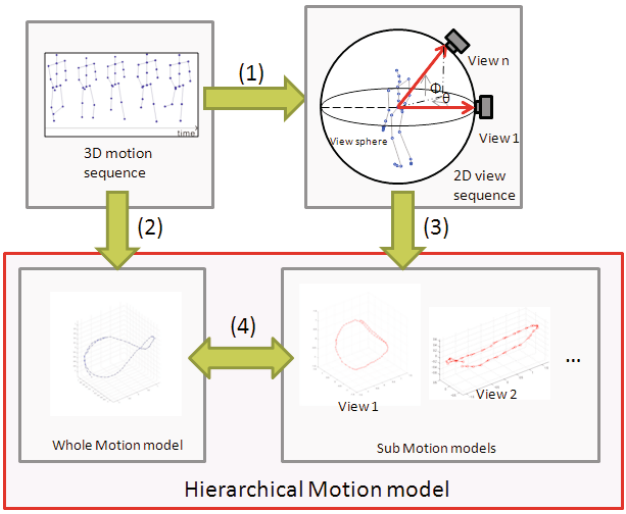


Fig. 2. Learning overview. (1) Use GPDM to map 3D human motion data to latent space. We call this the "whole motion model". (2) Map 3D human motion data to the view sphere surface. (3) Use GPDM to map 2D human motion data to latent space for each view data element. We call these the "sub motion models". (4) Use GPR to get regression functions from the sub motion models for the whole motion model. Collectively, we call the whole motion model, the sub motion models, and the regression functions the "hierarchical motion model".

$$\mathbf{x}^*_{all} = \mathbf{C}^* {\mathbf{C_N}}^{-1} \mathbf{X}_{all} \tag{9}$$

here, $\mathbf{C}^*$ denotes a kernel function, $\mathbf{x}^*_{sub}$ a estimated sub latent variable, and $\mathbf{x}_{sub}$ a learned sub latent variable.

4.2 Particle Filter

In what follows, we will describe the method of tracking the human motion on an actual video, utilizing the obtained motion models. Our method is based on the particle filter. However, the problems of how to distribute particles and how to propagate them are still open for our task. For the first problem, we adopt an approach which distribute the particles in multiple sub motion models.

In case of the approach which distribute the particles, in the whole motion model., to evaluate each particle, the camera angle must be given for projecting the 3D joints onto the 2D screen. In previous works, such camera angle estimation is considered as a pre-processing. Also, in many researches, the camera angle is considered as a given parameter. Both these approaches are not effective and limited the usabilities for real scenes. Since each sub motion model is learned individually and is related to each viewpoint, particles can be distributed in such multiple models. Also each particle can be evaluated directly on the screen

because the learning sources are 2D coordinates. Next, we will describe the problem of how to propagate these particles in such multiple motion models. In fact, the propagation can be divided into two steps, which are resampling and predicting. In our task, the predicting step is defined clearly because it is defined by the motion models themselves. So the propagation problem can be derived to the problem of how to resample the particles. As a particle filter, the resampling is done by calculating the next distribution of particles from the current likelihood of all particles. In our task, for the particles distributed within one single sub motion model, the resampling can reasonably be done as same as the naive one. However, the distribution over different sub motion model, which means different camera angle, must also be updated according to the likelihood of all particles. As a summary, by distributing particles on multiple sub motion models, the evaluation of each particle is achieved directly. By resampling the particles both within and over the involved sub motion models, particles can be propagated in multiple models. As a result, both motion tracking and camera angle estimation can be achieved simultaneously. Fig.3 is the estimation overview.

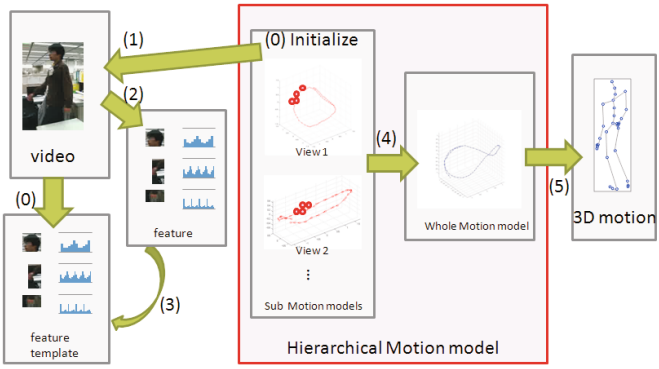


Fig. 3. Estimation overview. (0) Initialize 2D tracking position and view, then get the feature corresponding 2D tracking point as a feature template. (1) Map from particles in the sub motion model to the 2D position. (2) Get the features corresponding to each particle. (3) Compare (2) and the feature template, and get weights of each particle. (4) Use GPR to map from a selected particle in the sub motion model to the whole motion model. (5) Map from the state variable in the whole motion model to the 3D pose.

5 Experiments and Estimation Results

We use two types of motion models: the whole model and our hierarchical model (see Fig.2). The motion data is single walker motion capture data captured by Motion Analysis Corporation's Eagle Digital RealTime System. The RBF kernel function, which we adopted on the basis of preliminary experiment results, is used

as GPR in Eq.5,. We estimate video sequences of the same person as learning data under two conditions: without camera operation, which is accompanied by a slight relative view change and with free camera operation, which is accompanied by a big relative view change. In initialization, we manually supply three 2D points in the first frame: head, right hand, and right foot. We supply the same initial view parameter to both methods, marking the root position in all frames as the human center position.

Figure 4 shows estimation results without and with camera operation. The top sequence is the results obtained with the whole motion model and the bottom one is those obtained with our proposed (hierarchical) model. The red lines show estimated human motion. With the hierarchical model early convergence is obtained, making stable estimation possible even with free camera operation. One reason for this is that the search space is more limited in the hierarchical model than in the whole motion model, because the latent space in the former is connected to 2D joint positions while in the latter it is connected to 3D joint positions. Another reason is that while the whole motion model is not able to deal with relative view change, the hierarchical motion model is able to generate particles in multi-view subspaces and thus is able to estimate both view and motion at one time. However, with free camera operation, which is accompanied by a big relative view change, the estimated precision is not sufficiently high. Particles were produced every five degrees within the space of nine neighborhoods, and we believe this is because the search space was too small to catch the camera movement. Red lines show estimated human motion.

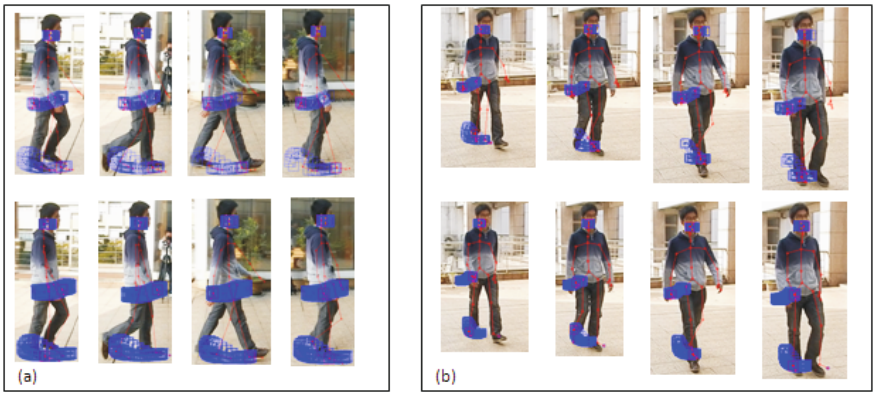


Fig. 4. Estimation results for every 20th frame. (a) is sequence without camera operation and (b) is sequence with free camera operation.The top sequence was obtained with the whole motion model and the bottom one with our proposed (hierarchical) model. The red line is estimated human motion. With the hierarchical model early convergence is obtained, making stable estimation possible even with free camera operation.

With the hierarchical model early convergence is obtained, making stable estimation possible even with free camera operation. One reason for this is that the search space is more limited in the hierarchical model than in the whole motion model, because the latent space in the former is connected to 2D joint positions while in the latter it is connected to 3D joint positions. However, with free camera operation, which is accompanied by a big relative view change, the estimated precision is not sufficiently high. Particles were produced every five degrees within the space of nine neighborhoods, and we believe this is because the search space was too small to catch the camera movement.

6 Conclusions

We have proposed a hierarchical motion model and a motion estimation method for it to estimate human motion with free camera setting and operation. We showed it can robustly estimate human motion for cases involving relative changes in human and camera viewpoints. In future work, there are two issues we will need to address. The first is how to develop an automatic motion category recognition method for estimating multiple target motions. And the second is incorporating a way to handle multi-modal input such as motions and sounds to improve accuracy of 3D motion estimation.

References

1. Moeslund, T.B., Hilton, A., Krüge, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. 104(2-3), 90–126 (2006)
2. Poppe, R.: Vision-based human motion analysis: An overview. 108(1-2), 4–18 (2007)
3. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-Time Human Pose Recognition in Parts from a Single Depth Image. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, pp. 1297–1304 (2011)
4. Shen, S., Tong, M., Deng, H., Liu, Y., Wu, X., Wakabayashi, K., Koike, H.: Model based human motion tracking using probability evolutionary algorithm. Pattern Recognition Letters 29(13), 1877–1886 (2008)
5. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian Process Dynamical Models. In: NIPS 2005. Adv. Neural Inform. Process. Systems, Vancouver (2005)
6. Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Machine Intell. 30(2), 283–298 (2008)
7. Urtasun, R., Fleet, D.J., Fua, P.: 3D People Tracking with Gaussian Process Dynamical Model. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, vol. 1, pp. 238–245 (2006)
8. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)

Motion History of Skeletal Volumes and Temporal Change in Bounding Volume Fusion for Human Action Recognition

Abubakrelsedik Karali and Mohammed ElHelw

Ubiquitous Computing Group, Centre for Informatics Sciences, Nile University, Egypt
a.karali@ieee.org, melhelw@nileuniversity.edu.eg

Abstract. Human action recognition is an important area of research in computer vision. Its applications include surveillance systems, patient monitoring, human-computer interaction, just to name a few. Numerous techniques have been developed to solve this problem in 2D and 3D spaces. However 3D imaging gained a lot of interest nowadays. In this paper we propose a novel view-independent action recognition algorithm based on fusion between a global feature and a graph based feature. We used the motion history of skeleton volumes; we compute a skeleton for each volume and a motion history for each action. Then, alignment is performed using cylindrical coordinates-based Fourier transform to form a feature vector. A dimension reduction step is subsequently applied using PCA and action classification is carried out by using Mahalonobis distance, and Linear Discernment analysis. The second feature is the temporal changes in bounding volume, volumes are aligned using PCA and each divided into sub volumes then temporal change in volume is calculated and classified using Logistic Model Trees. The fusion is done using majority vote. The proposed technique is evaluated on the benchmark IXMAS and i3DPost datasets where results of the fusion are compared against using each feature individually. Obtained results demonstrate that fusion improve the recognition accuracy over individual features and can be used to recognize human actions independent of view point and scale.

1 Introduction

Human motion analysis is a key area in computer vision research. Human motion analysis is divided into action recognition and activity recognition. The former typically deals with identifying actions each represented by short, and occasionally periodic, motion patterns such as walking, jumping, running, jogging, .etc. In the latter, long and complex motion patterns are used to identify human activity and group interactions. This paper focuses on human action recognition systems where the goal is to automatically classify ongoing activities in unlabeled videos. The applications of human action recognition systems include surveillance systems, patient monitoring systems, and a variety of systems that involve human-computer interfaces. In the simple case where a video is segmented to contain only one human action, the objective of the system is to correctly classify the video into its action category.

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 79–87, 2013.

Action and activity recognition approaches are classified into two categories [1]: single-layered approaches and hierarchical approaches. Single-layered approaches are approaches that represent and recognize human actions directly based on sequences of images. On the other hand, hierarchical approaches recognize high-level human activities by describing them in terms of actions, which they generally call sub-events. Single-layered approaches are classified into two types depending on how they model human activities: space-time techniques and sequential techniques. Space-time techniques represent action videos in three-dimensions, 2D video images over time domain, while sequential techniques interpret an action video as a sequence of observations. Space-time techniques are further divided into three categories: space-time spaces themselves, trajectories, or local interest point descriptors.

The proposed approach belongs to space-time spaces techniques. Space-time approaches are first introduced by [2] it was scale invariant, however the algorithms was limited to cyclic actions. Then [3, 4] introduced a view based template approach using Motion Energy Images (MEI) and motion history images (MHI) to indicate presence of motion and the order respectively but this approach was view variant and limited to 2D cases. In [5, 6] they extends the MHI into the 3D space and introduced the Motion History Volumes (MHV), in [7] they extended the research done in MHV and introduced Motion history of skeletonized.

In this research, we extended our work [7] and propose fusion between motion history skeletal volumes (MHSVs) and temporal changes in bounding volumes (TCBV) in order to improve accuracy of the action recognition as MHSVs are sensitive to the errors in 3D reconstruction and this may affect the accuracy and motion history feature is not reliable for very long actions. The key contribution of the paper is the usage of multiple and diverse features with deferent classification techniques for improved human action recognition in videos. The paper is organized as follows: Section 3 provides necessary definitions. Section 4 describes the proposed work while Section 5 presents obtained results and discussion whereas future work is provided in Sections 5 and 6, respectively.

2 Methodology

The proposed technique is composed of extracting MHSV features and classification, temporal change in bounding volume (TCBV) feature extraction and classification and finally the fusion using majority vote, as illustrated in Figure 1.

Concerning MHSV there are two stages: (1) MHSV feature extraction, and (2) processing and classification. In the first stage, actions are input in the form of 3D binary blobs called visual hulls [8]. Visual hulls are then skeletonised and a motion history volume is subsequently computed for each action from the skeletonised samples to obtain MHSV features. In the second stage, MHSVs are processed by applying cylindrical coordinates Fourier transform to produce rotation-invariant feature vectors. Before classification, PCA is applied for dimension reduction by projecting the data along the principle axes that maximize the discrimination between samples. A single Gaussian model is used to represent each action where the model is described by the distribution mean and variance. Action classification is then carried out by using 3 different techniques: Mahalanobis distance, and Linear Discriminate

Analysis (LDA). Details of the proposed algorithm are next described. TCBV is (1) calculated from the volume after alignment and division in to 3 parts. (2)The outputs then preprocessed using PCA and then classified using LMT. Finally the output from each is then fused using majority rule.

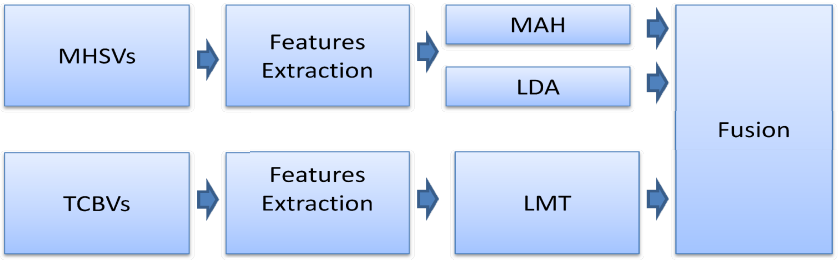


Fig. 1. Flow Diagram of the proposed action recognition algorithm

2.1 MHSVs

Feature Extraction

Skeletons are computed from three dimensional visual hull following the algorithm presented in [9] where a skeleton is defined as the singularities in the Euclidean distance-to-boundary field. When the distance transform is seen as a height map, the singularities can be seen as the local peaks of the distance transform. In order to extract skeletons, we calculate the divergence of the distance transform and a threshold is used to extract the skeleton and preserve the connectivity of the skeleton parts.

For each action we compute the motion history of the skeletal volume (MHSV) based on Equation 1. Generating 64*64*64 element volume, each motion history is then scaled in time and spatial domains and centered to the occupying volume for the next step. In order to get a rotation invariant feature, MHSVs are then transformed into cylindrical coordinates [6] in which a rotation is represented as translation using:

$$\mathrm{v}\left(\sqrt{\mathrm{x}^2+\mathrm{y}^2}, \tan\left(\frac{\mathrm{x}}{\mathrm{y}}\right), \mathrm{z}\right) \rightarrow \mathrm{v}(\mathrm{r}, \theta, \mathrm{z}), \tag{1}$$

where (x,y,z) and (r, θ, z) are the basis of the original coordinates and the cylindrical coordinates at [5], respectively. A 3D Fourier transformation is carried out using Equation 2.

$$V(k_r, k_\theta, k_z) = \int_{-\infty}^{\infty} \int_{-\pi}^{\pi} \int_{-\infty}^{\infty} V(r, \theta, z)\, e^{-j(K_r r + k_\theta \theta\, k_z z)} dr d\theta dz, \tag{2}$$

where k_r, k_θ, k_z are the bases of the Fourier coordinates. The output of Fourier transform is used as an action the feature vector for processing and classification.

Processing and Classification

A dimensional reduction using PCA is then applied to the feature vectors for dimension reduction and to project the data along the principle axes that maximize the discrimination between samples. A covariance matrix is first computed for *N* training samples using:

$$\sum(i,j) = \frac{1}{N}(f_i - \mu)\left(f_j - \mu\right)^{\mathrm{T}} \tag{3}$$

where μ represents the average of all feature vectors *f*, and the superscript *T* donates the transpose of the matrix. The covariance $\sum$ is then decomposed into the Eigen values and the corresponding Eigen vectors. The Eigen values are then sorted decreasingly and the Eigen vectors matrix is reformed according to the sorted Eigen values resulting into a projection matrix that is used to define a new projective space.

Each action is defined by a single Gaussian model represented by the mean μ and variance σ over all training set for the same action. All testing samples are then projected into the new projective space using the projection matrix and assigned to the class of minimum distance based on:

Mahalanobis distance, given by:

$$d_m(\mu_j, f) = \sqrt{\sum_i \frac{(\mu_{j_i} - f_i)^2}{\sigma_i}} \tag{4}$$

where μ_j is the mean of the model represent the class *j*, *f* is the test feature vector after projection, μ_i is the i^{th} element in the mean μ and *f*, and σ_i is the variance for this element over all training samples.

For further data reduction and better classification we used Linear Discriminant Analysis (LDA) [10]. According to Fisher discriminant analysis, the separation between classes is the ratio of the variance between classes to the variance within the classes. In our case, if *y* is the feature vector after PCA, μ_j is the mean of each action model and μ is the average of all samples, the variance matrix within class S_w and between classes S_b is given by:

$$S_w = \sum_{i=1}^{c} \sum_{j}^{ni} (y_i - \mu_i)(y_i - \mu_i)^T \tag{5}$$

$$S_b = \sum_{i=1}^{c} (\mu_j - \mu\ )(\mu_j - \mu\ )^T \tag{6}$$

where *c* is the number of classes and *ni* is the number of samples per class *i*. A projection matrix *W* that maximize S_b and minimize S_w would be equal to the largest Eigen values of $S_w^{-1} * S_b$ [5] and sample z is assigned to the class *j* that makes Equation 7 minimum.

$$d(\alpha_j, z) = \left\|\alpha_j - z\right\|^2 \tag{7}$$

Table 1. Comparison between the accuracy of each individual classifier and the fusion model for i3DPost dataset

I3DPOST	Walk	Run	Jump forward	Jump in place	Bend	Wave	Sit/stand	Walk/sit	Run/fall	Run/ fall/jump	Hand shaking	Average
MHSV/LDA	**0.8**	**0.8**	**1**	**0.6**	**0.8**	**1**	**1**	**0.8**	**0.6**	**0.8**	**1**	**0.84**
MHSV/MAH	**0.6**	**0.8**	**0.8**	**0.6**	**0.8**	**1**	**1**	**0.6**	**0.6**	**0.8**	**1**	**0.78**
TCBV/LMT	**0.77**	**0.83**	**0.8**	**0.97**	**0.97**	**0.73**	**0.93**	**0.73**	**0.63**	**0.67**	**0.67**	**0.78**
Fusion	0.8	1	0.4	1	1	1	1	0.6	0.8	1	1	0.89

2.2 TCBVs

Feature Extraction

Staring from 3D volumes we got all volumes aligned with the same orientation using PCA of the spatial information we got all volumes aligned with the same orientation using PCA of the spatial information. Then, we divide each volume in to three sub-volumes along z-direction which is described by the first principle axis represents the vertical. These segments are head, body and legs following the human body ratio. We generate a temporal curve represent the bounding box x and y size for each sub volume and a global z to generate seven temporal curves.

Processing and Classification

Each curve is then normalized by its mean value and resized to be 100 samples in temporal resolution. The curves are then concatenated to form one feature vector of length 700. PCA is then applied to this feature vector for data reduction and to find the axes that most discriminate the feature vectors. The result feature vector is then input to the classification stage. We use logistic model trees (LMT) for classification which are classification trees with logistic regression functions at the leaves.

2.3 Fusion

We implemented our approach using against different fusion role. We used majority rule; it is a decision rule that selects alternatives which have a majority of votes. For MHSVs we used the first two predictions from both Mahalanobis distance and LDA together with the prediction of the LMT using TCBV. For further evaluation we used Logistic model trees[11], Decision trees C4.5 implementation[12], K-Star[13], Multilayer Perceptron and Naïve Bayesian rule.

3 Results and Discussion

We evaluate our approach using two benchmark datasets the IXMAS and the i3DPost. Details of each dataset are as following:

Table 2. Comparison between the accuracy of each individual classifier and the fusion model for IXMAS dataset

IXMAS	Check Watch	Cross Arms	Scratch Head	Sit down	Get up	Turn Around	Walk	Wave	Punch	Kick	Point	Pickup	Average
MHSV/LDA	0.7	0.8	0.63	1	1	0.97	1	0.63	0.67	0.9	0.67	0.83	0.82
MHSV/MAH	0.7	0.8	0.73	1	0.97	0.97	1	0.77	0.67	0.93	0.63	0.87	0.84
TCBV/LMT	0.77	0.83	0.8	0.97	0.97	0.73	0.93	0.73	0.63	0.67	0.67	0.8	0.79
Fusion	**0.77**	**0.87**	**0.73**	**1**	**0.97**	**0.97**	**1**	**0.77**	**0.7**	**0.87**	**0.73**	**0.93**	**0.88**

INRIA's IXMAS Motion Acquisition Sequences dataset **[14]** we used 12 actions. each performed 3 times by 10 actors (5 males / 5 females). The acquisition was achieved using 5 cameras. To demonstrate view-invariance, the actors freely changed their orientation for each acquisition. Silhouettes were extracted from the videos using standard background subtraction techniques. Afterwards, visual hulls are carved from a discrete space of voxels of resolution of 64*64*64.

The I3DPost dataset [15] is an action database that has been created by using eight camera setup to produce multi-view videos. Each video depicts one of eight persons (2 females and 6 males) performing one of twelve different human motions, six actions and six activities. The subjects have different body sizes, clothing and are of different sex and nationality. The database contains 104 multi-view videos or 832 (8 × 104) single-view videos. The multi-view videos have been further processed to produce a 3D mesh at each frame describing the respective 3D human body surface. We voxelized the 3D human meshes using mesh rasterization followed by hole filling and then resized it into 64*64*64 volume.

The performance of the proposed fusion model is compared against the individual features of MHSV and TCBV in order to show the improvement in reuslts. For each dataset, action classification is carried out by using LDA, and Mahalanobis distance over all actors. The average accuracy of classification is computed and a pooled con fusion matrix is subsequently created for each classification technique. A leave-one-out routine is used to split action data into learning and testing where one actor is iteratively selected out of the training data and used for testing. As shown in tables 1, the action recognition accuracy achieved by applying the proposed MHSV approach on the I3DPost dataset is 83.6%, 78.1% for the LDA, and Mahalanobis distance using MHSV, 78.18% using TCBV with LMT and **89.1%** for fusion using majority vote, respectively.

For the IXMAS dataset, the action recognition accuraciesare 81.67%, 83.61% for the LDA, and Mahalanobis distance using MHSV, 79.17% using TCBV with LMT and **86.61%** after fusion. The class accuracies of the classes are shown in tables 1 and 2 and, table 1 and 2 illustrate the performance of the three action classification techniques when MHSV with LDA and Mahalanobis distance, TCBV with LMT and are used with the I3DPost and the IXMAS datasets, respectively. Table 4 shows the accuracy of the proposed approach against the corresponding accuracies of the state

Table 3. The results of desicion fusion using different fusion roles

%	LMT	DT	Kstar	MLP	NB	vote
I3DPOST	73.942	85.456	79.699	63.033	75.457	89.1
IXMAS	56.666	78.332	70	45	63.332	88.48

Table 4. Comparison between our approach against the state of arts aproachs on IXMAS and i3DPost Datasets

IXMAS		Number of actions	
Year	**Author**	**11**	**12**
2006	Weinland [5]	93.91%	79.72%*
2007	Weinland[17]	81.27%	
2007	Lv [18]		80%
2008	Turaga [16]	98.33%	
2008	Yan [19]	78%	
2011	Liu [20]		82.8%
2012	Karali [7]		83.6%
2012	**Our**		**88.48%**

I3DPost		Number of actions			
Year	Author	4	8	10	11
2009	Gkalelis [21]	90%			
2010	Iosifidis [22]		91%		
2011	Holte [23]	84.4%		80%	
2012	Karali[7]				84%
2012	**Our**				**89%**

*We repeat their Experiment on the ground truth sequences using 12 actions

of art approaches on IXMAS and I3DPostdatasets. In [5, 16, 17], it should be noted that only subsequences that maximally represent the action were used in their experiments, and these subsequences were selected manually. These subsequences are generated by splitting each action into sub actions using the change in motion energy. However higher accuracies they achieved, it cannot be generalized because of the lack of automation. From the tables we can see that the accuracy drops by increasing the number of actions.

As demonstrated in the above results, the proposed fusion model for action recognition approach offer improved recognition accuracy over the individual features when used on two of the benchmark datasets. Moreover according to i3DPost results it is suitable for recognition, of not only short actions, but also longer actions and short activities.

4 Conclusion

In this paper we discussed the fusion of MHSV and TCBV for human action classification. First, we compute the skeleton for each volume, then a motion history for each action. Then alignment is performed using cylindrical co ordinates based Fourier Transform forming feature vector. A dimension reduction step is then applied using Principle Component Analysis and finally classification is performed by using Mahalonobis distance and Linear Discernment analysis. Then we divided the volume into main three sub volumes and extracted the change in the bounding volume as spatio temporal feature and used LMT for classification. Finally we fuse the output of the classifiers using majority vote. The proposed algorithm is evaluated on the benchmark IXMAS and i3DPost datasets where the proposed approach is compared

against the each single feature and against the state of arts. Obtained results demonstrate that skeleton representations improve the recognition accuracy and can be used to recognize human actions independent of view point and scale.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Comput. Surv. 43(3), 1–43 (2011)
2. Polana, R., Nelson, R.: Recognizing activities. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference A: Computer Vision & Image Processing, vol. 1 (1994)
3. 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)
4. Davis, J.W.: Hierarchical motion history images for recognizing human motion. In: Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video 2001 (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. Weinland, D., Ronfard, R., Boyer, E.: Motion history volumes for free viewpoint action recognition. In: IEEE International Workshop on Modeling People and Human Interaction, vol. 104(2) (2005)
7. Karali, A., ElHelw, M.: Motion History of Skeletal Volumes for Human Action Recognition. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Fowlkes, C., Wang, S., Choi, M.-H., Mantler, S., Schulze, J., Acevedo, D., Mueller, K., Papka, M. (eds.) ISVC 2012, Part II. LNCS, vol. 7432, pp. 135–144. Springer, Heidelberg (2012)
8. Laurentini, A.: The Visual Hull Concept for Silhouette-Based Image Understanding. IEEE Trans. Pattern Anal. Mach. Intell. 16(2), 150–162 (1994)
9. Bouix, S., Siddiqi, K.: Divergence-Based Medial Surfaces. In: Vernon, D. (ed.) ECCV 2000 Part I. LNCS, vol. 1842, pp. 603–618. Springer, Heidelberg (2000)
10. Swets, D.L., Weng, J.J.: Using discriminant eigenfeatures for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(8), 831–836 (1996)
11. Landwehr, N., Hall, M., Frank, E.: Logistic Model Trees. Mach. Learn. 59(1-2), 161–205 (2005)
12. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc. (1993)
13. Cleary, J.G., Trigg, L.E.: K*: An Instance-based Learner Using an Entropic Distance Measure. In: Proceedings of the 12th International Conference on Machine Learning, pp. 108-114 Key (1995)
14. 4D Repository, INRIA Xmas Motion Acquisition Sequences (IXMAS). INRIA Xmas Motion Acquisition Sequences (IXMAS), `http://4drepository.inrialpes.fr/public/viewgroup/6` (cited 2012)
15. Gkalelis, N., et al.: The i3DPost Multi-View and 3D Human Action/Interaction Database. In: Proceedings of the 2009 Conference for Visual Media Production, pp. 159–168. IEEE Computer Society (2009)
16. Turaga, P., Veeraraghavan, A., Chellappa, R.: Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008 (2008)

17. Weinland, D., Boyer, E., Ronfard, R.: Action Recognition from Arbitrary Views using 3D Exemplars. In: IEEE 11th International Conference on Computer Vision, ICCV 2007 (2007)
18. Fengjun, L., Nevatia, R.: Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007 (2007)
19. Pingkun, Y., Khan, S.M., Shah, M.: Learning 4D action feature models for arbitrary view action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008 (2008)
20. Liu, J., et al.: Cross-view action recognition via view knowledge transfer. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3209–3216. IEEE Computer Society (2011)
21. Gkalelis, N., Nikolaidis, N., Pitas, I.: View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation. In: Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, pp. 394–397. IEEE Press, New York (2009)
22. Iosifidis, A., Nikolaidis, N., Pitas, I.: Movement recognition exploiting multi-view information. In: 2010 IEEE International Workshop on Multimedia Signal Processing, MMSP (2010)
23. Holte, M.B., et al.: 3D Human Action Recognition for Multi-view Camera Systems. In: Proceedings of the 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, pp. 342–349. IEEE Computer Society (2011)

Multi-view Multi-modal Gait Based Human Identity Recognition from Surveillance Videos

Emdad Hossain, Girija Chetty, and Roland Goecke

Faculty of ISE,
University of Canberra, Australia
emdad.hossain@canberra.edu.au

Abstract. In this paper we propose a novel human-identification scheme from long range gait profiles in surveillance videos. We investigate the role of multi view gait images acquired from multiple cameras, the importance of infrared and visible range images in ascertaining identity, the impact of multimodal fusion, efficient subspace features and classifier methods, and the role of soft/secondary biometric (walking style) in enhancing the accuracy and robustness of the identification systems, Experimental evaluation of several subspace based gait feature extraction approaches (PCA/LDA) and learning classifier methods (NB/MLP/SVM/SMO) on different datasets from a publicly available gait database CASIA, show significant improvement in recognition accuracies with multimodal fusion of multi-view gait images from visible and infrared cameras acquired from video surveillance scenarios.

Keywords: multimodal, multiview, PCA, LDA, MLP, identification, SMO, feature fusion.

1 Introduction

Automatic human identification from arbitrary views is a very challenging problem, especially when one is walking at a distance. Over the last few years, recognizing identity from gait patterns has become a popular area of research in biometrics and computer vision, and one of the most successful applications of image analysis and understanding. Also, gait recognition is being considered as a next-generation recognition technology, with applicability to many civilian and high security environments such as airports, banks, military bases, car parks, railway stations etc. For these application scenarios, it is not possible to capture the frontal face, and even if it can be captured, it is of low resolution. Hence most of traditional approaches used for face recognition fail. However, several physiological and biomechanical studies have shown that human gait is inherently multimodal, and is based on kinematic interaction between several motion articulators, such as lower and upper limbs and other biomechanics of joints. It is person specific based on body weight, height, joint mobility in the limbs, and other behavioural nuances. If we can model these inherently multimodal traits , it is possible to identify human from a distance from their gait or from the way they walk. Even if frontal face is not visible, it is possible to establish the

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 88–99, 2013.

identity of the person using certain static and dynamic multimodal cues from frontal and profile face, ear and head shape, walking style and speed, hand motion during walking etc. If automatic identification systems can be built based on this concept, it will be a great contribution to surveillance and security area Further. this will make a significant contribution to better understanding of gait abnormalities, and development of human computer interfaces. However, each of these cues or traits captured from long range low resolution surveillance videos on its own are not powerful enough for ascertaining identity, A combination or fusion of each of them, along with an automatic processing technique can result in robust recognition. In this paper, we propose usage of full profile silhouettes of persons from visible range and infrared range camera footage for capturing inherent multi-modal cues available from of the gait patterns of the walking human. This also addresses the need to establish identity from low resolution surveillance video images. In addition, user cooperation is not mandatory upon data collection making it suitable for surveillance and law enforcement scenarios. Further, capture of long range gait biometric data from surveillance videos contains several biometric traits such as side face, ear, body shape, and gait motion, which are a combination of physiological and behavioural biometrics. Automatic schemes that can process this rich multimodal information can result in robust human identification approaches.

In this paper, we propose the use of a principled approach involving feature extraction techniques based on multivariate statistical techniques, such as principle component analysis (PCA) and linear discriminant analysis (LDA), and efficient learning classifier approaches based on support vector machines and Bayesian classifiers. Further, we propose that the feature level fusion of multi-view multispectral images (from visible range cameras and infrared cameras) can enhance the performance of identification scheme as compared to single mode image features. Fusing features at the feature level is more effective than fusion at later stages, as the inherent multi-modality can be preserved at early stages of processing as compared to late fusion [2]. The experimental evaluation of the proposed approach with a publicly available CASIA [1] gait database shows a significant improvement in recognition performance as compared to other methods proposed in the literature. Rest of the paper is organised as follows. Next Section describes the background and motivation for proposed work, followed by the proposed multiview multimodal identification scheme in Section 3. The details of the experiments performed is described in Section 4, and conclusions and plans for further work is described in Section 5.

2 Background

Current state-of-the-art video surveillance systems, when used for recognizing the identity of the person in the scene, cannot perform very well due to low quality video or inappropriate processing techniques. Though much progress has been made in the past decade on visual based automatic person identification through utilizing different biometrics, including face recognition, iris and fingerprint recognition, each of these techniques work satisfactorily in highly controlled operating environments such as border control or immigration check points, under constrained illumination, pose and facial expression variations. To address the next generation security and surveillance

requirements for not just high security environments, but also day-to-day civilian access control applications, we need a robust and invariant biometric trait [3] to identify a person for both controlled and uncontrolled operational environments. According to authors in [4], the expectations of next generation identity verification involve addressing issues related to application requirements, user concern and integration. Some of the suggestions made to address these issues were use of non-intrusive biometric traits, role of soft biometrics or dominant primary and non-dominant secondary identifiers and importance of novel automatic processing techniques. To conform to these recommendations; often there is a need to combine multiple physiological and behavioral biometric cues, leading to so called multimodal biometric identification system.

Each of the traits, physiological or behavioral have distinct advantages, for example; the behavioral biometrics can be collected non-obtrusively or even without the knowledge of the user. Behavioral data often does not require any special hardware (other than low cost off the shelf surveillance camera). While most behavioral biometrics are not unique enough to provide reliable human identification they have been proved to be sufficiently high accurate [5, 6]. Gait, is a powerful behavioral biometric, but as a single mode, on its own it cannot be considered as a strong biometric to identify a person. However, if we combine complementary gait information from another source, the multi-modal combination is expected to be powerful for human identification. Researchers have found that one of the most promising techniques is the use of multimodality or combination of differnt biometric traits or same biometric trait from multiple disparate sources. For example, researchers in [7, 8] have found that multi-modal scheme involving PCA on combined image of ear and face biometric results in significant improvement over either individual biometric. In addition, other recent attempts to improve the recognition accuracy include face, fingerprint and hand geometry [9]; face, fingerprint and speech [10]; face and iris [11]; face and ear [12]; and face and speech [13]. The fusion of complementary biometric information from disparate sources, however, did not attract much attention from the research community. This could be due to difficulty in acquiring the data, and processing and making sense out of them.

3 Multimodal Identification Scheme

For experimental evaluation of our proposed multimodal gait identification scheme, we used CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences [1]. It is a large multi-view gait database, which is created in January 2005. There are more than 300 subjects. We used two different set of data known as dataset B and Dataset C. Dataset B was captured from 11 views with normal video camera, and 11 different views know as view angles. We used the data captured only in 90 degree view angle. The dataset C was captured with an infrared (thermal) camera. It takes into account four walking conditions: normal walking, slow walking, fast walking, and normal walking with a bag. The videos were all captured at night. Figure 1 shows the sample images in different view angles.

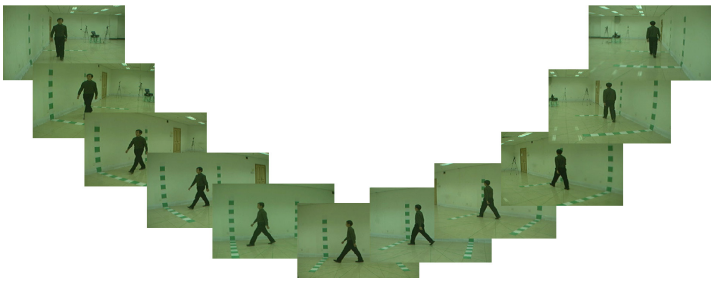


Fig. 1. Sample images from CASIA gait database

However, we used 50 subjects with a set of extracted silhouettes from Dataset B and another set of extracted silhouettes from Dataset C. Each subject consists of 16 images and in total 1600 images for 100 subjects (people). Figure 2 shows the extracted silhouettes from dataset B and C.

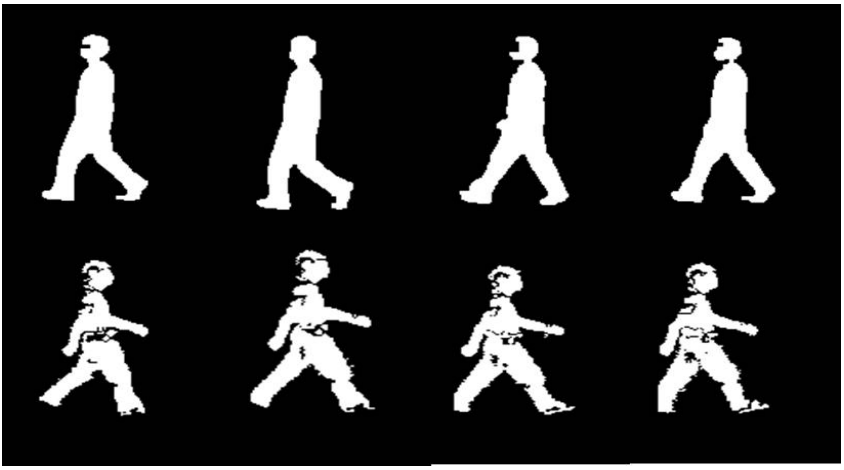


Fig. 2. Extracted silhouettes

We extracted the reduced dimensionality feature vector for each of the dataset separately by suing PCA (principal component analysis) and Linear Discriminant Analysis (LDA), and then have classified with different learning classifiers. Therefore our (cross camera feature level fusion) experiments involved evaluation of diiferent feature extraction and learning classifier combinations including PCA-MLP, LDA-MLP, PCA-SMO, and LDA-SMO.

3.1 Feature Extraction Using PCA-LDA Approach

Principle component analysis is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data. The other

main advantage of PCA is that once we have found these patterns in the data, and we can compress the data, e.g. by reducing the number of dimensions, without much loss of information. Basically this technique used in image compression [14]. In the image analysis it works like;

$$X=(x1, x2, x3\ldots\ldots N2) \quad (1)$$

where the rows of pixels in the image are placed one after the other to form a one dimensional image. Each image is *N* pixels high by *N* pixels wide. For each image it creates an image vector. And then it counts all the images together in one big image-matrix like;

$$Matrix = (v1, v2, v3\ldots\ldots vN) \quad (2)$$

On the other hand, the LDA also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made. LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis [15]. And in our experiment, LDA shows prominent than PCA. Next Section describes several classifiers we examined.

3.2 Naive Bayes and MLP Neural Network Classifier

Naive Bayes classifier can serve as a baseline classifier due to its simple probabilistic nature based on applying Bayes' theorem with strong (naive) independence assumptions. In other words, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without using any Bayesian methods [23]. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. Multi Layer perceptron (MLP) is a feedforward neural network with one or more layers between input and output layer. Feedforward implies that the data flows in on direction from input to output layer (forward). This type of network is trained with the backpropagation learning algorithm. MLPs are widely used for pattern classification, recognition, prediction and approximation. Multi Layer Perceptron can solve problems which are not linearly separable [16].

3.3 SVM and SMO Classifiers

Support Vector Machine (SVM) classifiers perform classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, in SVM, the original finite-dimensional space is mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function selected to suit the problem.[24] The hyperplanes in the higher-dimensional space are defined as the set of points whose inner product with a vector in that space is constant.

SMO, on the other hand is an SVM classifier with learning based on Sequential Minimal Optimization (SMO). SMO decomposes the overall QP problem into QP sub-problems, using Osuna's theorem to ensure convergence [16]. Unlike the other methods, SMO chooses to solve the smallest possible optimization problem at every step. The advantage of SMO lies in the fact that solving for multi instance multipliers can be done analytically. In addition, SMO requires no extra matrix storage at all. There are two components to SMO: an analytic method for solving for the two Lagrange multipliers, and a heuristic for choosing which multipliers to optimize [17].

$$y1 /= y2 \Rightarrow \alpha1 - \alpha2 = k \quad (1)$$

$$y1 = y2 \Rightarrow \alpha1 + \alpha2 = k \quad (2)$$

However, the multi instance multipliers must fulfil all of the constraints of the full problem. The linear equality constraint causes them to lie on a diagonal line. Therefore, one step of SMO must find an optimum of the objective function on a diagonal line segment [17].

4 Experiments and Results

Different sets of experiments were performed on two datasets in CASIA database-Dataset B containing visible normal images of walking humans, and Dataset C consisting of infrared images. By using PCA and LDA techniques, we extracted the feature vector for both datasets, training different learning classifiers and performed

identification experiments with multiple fold cross validation in single mode and multimodal fusion mode. We used different combinations of features (for example PCA-Dataset B, PCA-Dataset C, LDA-Dataset B, LDA-Dataset C and the feature level fusion of visible and infrared gait images from Dataset B and Dataset C. Table 1 to Table 5 show the recognition performance for each set of experiments in terms of recognition accuracy and several statistically significant performance measures such as true positive rate (TPR), false positive rate (FPR), precision, recall and Fmeasure.

All experiments involved either 5 or 10 fold cross validation. Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. One fold of cross-validation involves partitioning a sample of data into complementary subsets (training and testing subsets), performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple folds of cross-validation are performed using different partitions, and the validation results are averaged over the folds. We examined 5 fold and 10 fold cross-validation for each set of experiments.

Table 1. Classifier Performance for Dataset B (Visible range dataset) with PCA features with 50 dimensions. (NB – naïve Bayes; MLP – Multilayer Perceptron; TPR-true positive rate; FPR – false positive rate).

Classifier	Folds	Accuracy(%)	TPR	FPR	Precision	Recall	F-measure
NB	10	48.63	0.49	0.01	0.49	0.49	0.48
NB	5	47.68	0.48	0.01	0.49	0.48	0.48
MLP	10	79.5	0.8	0	0.8	0.8	0.79
MLP	5	75.13	0.75	0.01	0.76	0.75	0.75

The first set of experiments involve Dataset B (visible range dataset) with 50 dimensional PCA features. As can be seen in Table 1, The recognition accuracy for naïve Bayes classifier with different 10-fold and 5 fold cross-validation is low, with 48.63 % for 10 folds and 47.68 for 5 folds. Using MLP neural net classifier (with backpropagation learning) results in better accuracy with 79.5% for 10 folds and 75.13% for 5 folds. However, the MLP classifier is computational intensive with long train and test times. This could be due to inability of PCA features to discriminate multiple classes (50 classes here) with the available data size or the structure of the neural network used.

The second set of experiments involved use of linear discriminant analysis features and use of support vector machine classifier. As can be seen in Table 2, the naïve Bayes classifier with 50 dimensional LDA features results in significant improvement in performance with 92.5% recognition accuracy as compared to 48.6 % with PCA features for 10 fold cross-validation (CV). With 5 fold CV, the LDA features result in an accuracy of 92.25% as compared to 47.68% for PCA features. Due to computational intensive nature of neural net classifiers, we examined SVM classifier

Table 2. Classifier Performance for Dataset B (Visible range dataset) with LDA features with 50 dimensions. (NB – naïve Bayes; MLP – Multilayer Perceptron; SVM-L(Support Vector Machine-Linear Kernel); SVM-RBF (Radial Basis Function Kernel); SVM-poly (Polynomial Kernel); SVM-Signmoid (Signmoidal kernel).

Classifier	Folds	Accuracy(%)	TP	FP	Precision	Recall	F-measure
NB	10	92.5	0.93	0	0.93	0.93	0.93
NB	5	92.25	0.92	0	0.93	0.92	0.92
SVM -L	5	81.13	0.81	0	0.82	0.81	0.81
SVM -L	10	78.75	0.78	0	0.81	0.79	0.78
SVM -RBF	5	31.30%	0.3	0.02	0.74	0.3	0.39
SVM -poly	5	27.63%	0.28	0.02	0.75	0.28	0.36
SVM -sigmoid	5	29.13%	0.29	0.02	0.74	0.29	0.38

for this set of experiments, as SVMs are known to have better generalization ability, are less computation intensive, and are based on sound theory, unlike neural networks whose development has followed a more heuristic path. Other advantages of SVM over neural networks are - whilst ANNs can suffer from multiple local minima, the solution to an SVM is global and unique, and SVMs have a simple geometric interpretation and give a sparse solution. Unlike ANNs, the computational complexity of SVMs does not depend on the dimensionality of the input space. ANNs use empirical risk minimization, whilst SVMs use structural risk minimization. SVMs outperform ANNs often, as they are less prone to overfitting [17]. However, the performance depends on the kernel used and other SVM parameters. As can be in Table 2, different types of kernels - linear kernel (SVM-L), radial basis function kernel (SVM-RBF), polynomial kernel (SVM-poly) and sigmoidal kernel (SVMsigmoid), result in different recognition accuracies. The SVM with linear kernel performs best with 81.3% recognition accuracy for 5 fold CV, and has a 78.75% for 10 fold CV. Also, for both naïve Bayes and SVM classifier with linear kernel, the performance with 5 fold cross-validation partition was almost similar to 10 fold cross validation. Hence, for rest of the experiments, we used 5 fold CV partition.

Table 3. Classifier Performance for Dataset C (Infrared range dataset) with LDA features with 5 folds. (NB – naïve Bayes; MLP – Multilayer Perceptron; SVM-L(Support Vector Machine-Linear Kernel); SMO(Poly)-Sequential Minimum Optimization-Polynomial Kernel.

Classifier	Features	Dim	Accuracy(%)	TPR	FPR	Precision	Recall	F-measure
NB	PCA	50	56.63%	0.57	0.01	0.59	0.57	0.57
SVM-L	PCA	50	79.88	0.8	0	0.81	799	799
SVM-L	LDA	50	86.25%	0.86	0	0.88	0.86	0.87
NB	LDA	50	93.75%	0.94	0	0.94	0.94	0.94
NB	LDA	25	93.5	0.94	0	0.94	0.94	0.94
SVM-L	LDA	25	83.25%	0.83	0	0.85	0.83	0.84
SMO -poly	LDA	25	94	0.94	0	0.95	0.94	0.94

For the third set of experiments, we examined Dataset C, the infrared camera gait image dataset, with 5 fold cross validation. As can be seen in Table 3. infrared image dataset performs better than visible range dataset for both PCA and LDA features. The recognition accuracy achieved with 50 dimensional PCA features results is 56.3% for naïve Bayes classifier for Dataset C as compared to 47.68% for Dataset B (Table 1). A similar improvement in performance was achieved with 50-dimensional LDA features resulting in a recognition accuracy of 93.75% for Dataset C as compared to 92.25% for Dataset B. Further, we also examined reduced dimensional LDA features, as LDA features seem to model the identities better, even with large number of classes (50 classes/subjects). As can be seen in Table 3, there is no significant loss of accuracy with reduced dimensional feature vectors. With 25 dimensional LDA feature vector, the recognition accuracy achieved was 93.5 % for naïve Bayes classifier (as compared to 93.75% for 50 dimensions) and the accuracies were 83.25% for SVM with linear kernel (86.25%). This has a significant advantage as the reduced dimensional feature vector results in improvement in computational speed. In addition, for this set of experiments, we examined a different version of SVM classifier – SMO, the SVM with Sequential minimal optimization(SMO). SMO classifier uses an efficient algorithm for solving the optimization problem needed for training of support vector machines, and is known to result in a better performance than a traditional SVM which uses much more complex quadratic optimization problem during training. As can be seen in Table 3, the recognition accuracy achieved with SMO classifier with polynomial kernel is 94% as compared to 93.25 % achieved with SVM classifier with linear kernel.

Table 4. Classifier Performance for fusion of visible and infrared gait images (Dataset B + Dataset C) with equal weights a and with LDA features with 5 fold cross validation. (NB - naïve Bayes; SVM-L(Support Vector Machine-Linear Kernel; SMO(Poly)-Sequential Minimum Optimization- Polynomial Kernel.

Classifier	Dim	Accuracy(%)	TPR	FPR	Precision	Recall	F-measure
NB	50	98.38%	0.98	0	0.99	0.98	0.98
SVM-L	50	74.88%	0.79	0.01	0.75	0.75	0.97
SMO-poly	50	98.25%	0.98	0	0.98	0.98	0.98
NB	25	98.50%	0.99	0	0.99	0.99	1
SVM-L	25	72.50%	0.73	0.01	0.77	0.73	0.73
SMO-Poly	25	97.75%	0.98	0	0.98	0.98	0.98

The fourth set of experiments involved the feature level fusion of visible and infrared images from Dataset B and Dataset C. As we found the LDA features to be more discriminatory as compared to PCA, we used LDA features for all fusion experiments. As can be seen in Table 4, the fusion of normal visible camera and infra red camera images is synergistic, resulting in improvement in recognition performance as compared to single mode images. For naïve Bayes classifier, 50-dimensional LDA features result in 98.38% accuracy and 25-dimensional LDA

features result in 98.5%. The recognition accuracy achieved with SVM-L (linear kernel) for 50-dim LDA features is 74.88% and 72.5% for 25-dim LDA vector. The SMO version of SVM classifier with polynomial kernel results in 98.25% accuracy for 50-dim LDA vector, and for 25 dimensional LDA features, the accuracy is 97.75%. Once again for fusion mode, SMO with polynomial kernel performs better than traditional SVM with linear kernel. An interesting observation was that the multimodal fusion (feature level) performs a more dominant role as compared to the type of classifier or the type of features, as irrespective of classifier used (naïve Bayes or SVM), the recognition accuracy is significantly higher with multimodal fusion (higher than 95 %).

The final set of experiments involved investigating the role of soft or secondary biometric information, in terms of walking style(fast walking and normal walking) for enhancing the recognition accuracy. The walking style data was available for visible camera images only for all 50 subjects (persons). We used the data for each person walking in two (2) different styles - fast and normal walking. In this final set of experiments, we examined three different approaches. First, we applied LDA-MLP separately to (1) normal walking data, (2) the fast walking data and (3) combined the data corresponding to slow and fast walking information into a single dataset. This represents a challenging scenario with both dominant identity specific gait information (primary biometric) and non-dominant secondary/soft biometric information (walking style) modeled by LDA/MLP approach.

Table 5. Result in fast walking and normal walking

No	Method	Dataset	Accuracy
1	LDA-MLP	Normal Walking	95.5%
2	LDA-MLP	Fast walking	94.5%
3	LDA-MLP	Combined	82.50%

As can be seen in Table 5, while individually fast and slow walking style information modeled by LDA/MLP technique results in good identification accuracy, with 95.5% for normal walking, and 94.5% for fast walking, the modeling of weak soft biometric information (walking style) along with strong biometric information (identity of each subject) weakens the overall identification accuracy (82.5%). However, this depicts more real world scenario, and development of appropriate high performance subspace features and efficient classifier methods can result in better identification performance. It should be noted that the fusion of primary and soft/secondary biometric features is not reported in Table 5 due to lack of space, but some of our preliminary experiments show that fusion of primary and secondary/soft biometric information (walking style) can result in synergistic fusion. Also, use of motion based static and dynamic features is currently being investigated.

5 Conclusions and Further Plan

In this paper we proposed a novel human-identification scheme from long range gait profiles in surveillance videos. We investigated the role of multi view gait images acquired from multiple cameras - infrared and normal visible images in ascertaining identity. We also examined the benefits achieved with multimodal fusion, the roles of efficient subspace features and classifier methods, and the importance of soft/secondary biometric (walking style) in enhancing the accuracy and robustness of gait based identification systems, Experimental evaluation of several subspace based gait feature extraction approaches (PCA/LDA) and classifier methods (NB/MLP/SVM/SMO) on different datasets from a publicly available CASIA gait database, showed a significant improvement in recognition accuracies with multimodal fusion of multiview gait images acquired from normal visible and infrared video surveillance scenarios.

References

1. Zheng, S.: CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences, CASIA Gait Database, http://www.sinobiometrics.com
2. Huang, L.: Person Recognition By Feature Fusion, Dept. of Engineering Technology Metropolitan State College of Denver. IEEE, Denver (2011)
3. Bringer, J., Chabanne, H.: Biometric Identification Paradigm Towards Privacy and Confidentiality Protection. In: Nichols, E.R. (ed.) Biometric: Theory, Application and Issues, pp. 123–141 (2011)
4. Jain, A.K.: Next Generation Biometrics. Department of Computer Science & Engineering, Michigan State University, Department of Brain & Cognitive Engineering, Korea University (2009)
5. Yampolskiy, R.V., Govindaraja, V.: Taxonomy of Behavioral Biometrics. Behavioral Biometrics for Human Identification, 1–43 (2010)
6. Meraoumia, A., Chitroub, S., Bouridane, A.: Fusion of Finger-Knuckle-Print and Palmprint for an Efficient Multi-biometric System of Person Recognition. IEEE Communications Society Subject Matter Experts for Publication in the IEEE ICC (2011)
7. Berretti, S., Bimbo, A., Pala, P.: 3D face recognaition using isogeodesic stripes. IEEE Transaction on Pattern Analysis and Machine Intelligence 32(12) (2010)
8. Yuan, L., Mu, Z., Xu, Z.: Using Ear Biometrics for Personal Recognition,School of Information Engineering, Univ. of Science and Technology Beijing. Beijing 100083, http://yuanli64hotmail.com
9. Ross, A., Jain, A.K.: Information fusion in biometrics. Pattern Recognition Letters 24, 2115–2125 (2003)
10. Jain, A.K., Hong, L., Kulkarni, Y.: A multimodal biometric system using fingerprints, face and speech. In: 2nd. Int'l. Conf. AVBPA, vol. (10), pp. 182–187 (1999)
11. Wang, Y., Tan, T., Jain, A.K.: Combining Face and Iris Biometrics for Identity Verification. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 805–813. Springer, Heidelberg (2003)
12. Chang, K., et al.: Comparison and Combination of Ear and Face Images in Appearance-Based Biometrics. IEEE Trans. PAMI 25, 1160–1165 (2003)

13. Kittler, J., et al.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20, 226–239 (1998)
14. Smith, L.I.: A tutorial on Principal Components Analysis
15. Linear discriminant analysis, Wikipedia, http://www.wikipedia.org
16. Multi Layer Perceptron, http://www.neoroph.sourceforge.net
17. Platt, J.C.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Microsoft Research. Technical Report MSR-TR-98-14 (17) (1998), http://jplatt@microsoft.com
18. Shlizerman, I.K., Basri, R.: 3D Face Reconstruction from a Single Image Using a Single Reference Face Shape. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2) (2011)
19. Hossain, E., Chetty, G.: Multimodal Identity Verification Based on Learning Face and Gait Cues. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part III. LNCS, vol. 7064, pp. 1–8. Springer, Heidelberg (2011)
20. Chin, Y.J., Ong, T.S., Teoh, A.B.J., Goh, M.K.O.: Multimodal Biometrics based Bit Extraction Method for Template Security. In: Faculty of Information Science and Technology, Multimedia University, Malaysia, School of Electrical and Electronic Engineering, Yonsei University, IEEE, Seoul (2011)
21. Multilayer Perceptron Neural Networks, The Multilayer Perceptron Neural Network Model, http://www.dtreg.com

Using the Transferable Belief Model for Multimodal Input Fusion in Companion Systems

Felix Schüssel, Frank Honold, and Michael Weber

Institute of Media Informatics
Ulm University
89081 Ulm, Germany
{felix.schuessel,frank.honold,michael.weber}@uni-ulm.de

Abstract. Systems with multimodal interaction capabilities have gained a lot of attention in recent years. Especially so called companion systems that offer an adaptive, multimodal user interface show great promise for a natural human computer interaction. While more and more sophisticated sensors become available, current systems capable of accepting multimodal inputs (e.g. speech and gesture) still lack the robustness of input interpretation needed for companion systems. We demonstrate how evidential reasoning can be applied in the domain of graphical user interfaces in order to provide such reliability and robustness expected by users. For this purpose an existing approach using the Transferable Belief Model from the robotic domain is adapted and extended.

Keywords: HCI, Multimodal Fusion, Multimodal Interaction, Companion Systems, Evidential Reasoning.

1 Introduction

Companion Systems yield properties of multimodality, individuality, adaptability, availability, cooperativeness and trustworthiness [14]. One characteristic of such systems is their use of a multitude of sensors to gain information about the entire situation of the user, machine, and environment. Advances in processing power, sensory technology, and recognition techniques make it relatively easy to create systems that are able to detect user inputs from "natural" modalities like speech and gesture. But just equipping a system with sophisticated sensors is not enough. The information provided by these sensors need to be combined and interpreted by an intelligent fusion mechanism that allows a system to infer the original concept that was expressed by the user via the different modalities. Much work has been done on this topic, but still approaches lack the reliability and robustness needed for companion systems. In this article we present the use of evidential reasoning in terms of the Transferable Belief Model for the task of fusing user inputs from different modalities. Based on an approach initially suggested by Reddy and Basir in the robotic domain [10] our approach allows graphical user interfaces to be extended by natural ways of interaction in a robust and still flexible manner.

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 100–115, 2013.

The following related work section presents recent fusion approaches in the HCI domain and motivates the use of evidential reasoning. Section 3 describes the basic ideas behind evidential reasoning or more precisely, the Transferable Belief Model, and how it can be adapted for the task of fusing user inputs from different modalities. Section 4 then illustrates our approach of applying this method in the broader context of GUI-based applications. Section 5 finally summarizes our approach and provides an outlook on future work.

2 Related Work

Research describes fusion at different abstraction levels, namely *feature*, *decision*, and *hybrid* level fusion [11,5,1]. At feature level, distinguishable features of media streams (e.g. color-historgrams from a video stream), usually presented by numerical values are combined to form a larger feature vector that can be used to make a decision. At decision level, multiple of these decisions are combined to form a fused decision vector on which a semantically higher decision is made. When a system mixes both techniques, usually at different abstraction levels, it realizes a hybrid multimodal fusion. In the domain of HCI, usually some kind of decision or hybrid level fusion is applied, that combines incoming events from different modality sensors (e.g. voice and pointing gesture) to form a combined concept, like the selection of an object. The following gives a brief overview of current approaches and states their capabilities and limitations.

In the domain of human robot interaction, Holzapfel [6] uses an application independent parser of input events, and application specific rules to perform a hybrid level fusion. The input of each modality (speech and 3D pointing gestures) is parsed into an n-best list of typed feature structures called tokens (a form of structured attribute/value pairs), that are then passed on to the fusion. The fusion itself relies on two kinds of rules. Firstly, there are constraint rules that determine whether a subset of tokens can be merged, and secondly, there are creation rules that construct the result of a merge. While the approach can be quite powerful and is able to combine arbitrary inputs, it comes with the burden of creating application specific constraints and construction rules, that can become quite complex. Also it does not account for ambiguity of sensor inputs, but just selects the output that has the highest confidence in the n-best list. Typed feature structures (or something similar) that are merged during the fusion process have been used earlier by other researchers [7,2], although there was no possibility to explicitly define rules that manage combination.

The approach taken by Pfleger [9] is quite similar to that of Holzapfel, but uses the notion of dialog turns (called local turn context) as the basis of multimodal fusion. All unimodal events that belong to a dialog turn are stored in a working memory (WM) in terms of typed feature structures. All elements in the WM are assigned activation values (based on confidence scores of modality recognizers) that fade over time until a threshold is reached an the element is removed from the WM. Production rules that consist of condition-action sequences operate

on the WM and can itself change the WM in case they are fired. If multiple rules can potentially be fired, the one that has the highest score is executed. The score of an executed rule gets decreased, while the scores of not executed rules are increased. The enhanced flexibility compared to Holzapfel's approach comes with the burden of balancing the scores of the production rules to yield a consistent system behavior. Handling of ambiguous and conflicting sensor inputs is done by simply choosing the highest scoring production rule.

With HephaisTK, Dumas [3] presents a whole toolkit for the development of multimodal interfaces. It is build on software agents and includes not only agents for receiving modality inputs, and performing fusion, but also a dedicated dialog management component. Regarding the fusion itself, it is a purely decision level approach built upon rules consisting of triggers and actions. Relying on SMUIML [4], the synchronicity of events can be specified in detail (e.g. parallel or sequential triggers), allowing to combine complex input sequences. The major drawback of this approach is its assumption that all sensors are totally confident in their decisions, because it lacks any mechanisms to assign and evaluate confidence scores. This holds true for all systems that perform solely decision level fusion, as the necessary information remains within the sensors.

A different approach is presented by Reddy and Basir [10] based on set theory and the transferable belief model applied to the domain of human robot interaction. Evidential reasoning is used to combine evidences coming from sensory inputs to form extended concepts. The set of possible evidences and combined concepts is defined by a so-called *conceptual graph*. This way, not only probabilities can be represented as in the approaches mentioned above, but also ambiguous and conflicting sensory evidences are explicitly taken into account. This allows for a more informative approach that is able to disambiguate and reinforce multimodal inputs coming from different sensors. The superiority of this approach compared to traditional Dempster-Shafer theory and Bayesian approaches dealing with probabilities is demonstrated by measuring the entropy remaining in the system before and after fusion.

While the rule-based approaches described above are good at handling complex sequences of inputs by specifying temporal constraints, they tend to ignore the ambiguous nature of sensory inputs, or rely on simple n-best lists to make a decision. Concordant with Reddy and Basir we maintain that a major role of multimodal fusion, in addition to combination, lies in reinforcement and disambiguation of multimodal inputs. These tasks are either not fulfilled by rule based systems at all or, when dedicated rules can be specified, they lack a formally well defined mechanism. Taking their approach as a basis, we aim at deploying evidential reasoning in the broader context of GUI-based applications, as the most common way of human computer interaction. Before presenting our approach, the mathematical basics of the transferable belief model and the use of conceptual graphs as representation of possible interactions are given in the next section.

3 Fusion Based on Evidential Reasoning

In the following sections, the basic notions of evidential reasoning and the transferable belief model are briefly introduced and its basic applicability for fusing multimodal inputs is depicted.

3.1 Transferable Belief Model

The *transferable belief model* (TBM) and its basic concept, Dempster-Shafer's theory of evidential reasoning (DS-theory), is a generalization of probability theory, and quantifies beliefs of sensors in propositions (events). It was proposed by Philippe Smets in the early 1990's [12]. It differs from the classic probability, in such way, that a belief does not state the actual probability that an event happened, but only the confidence of the sensor about the event. The biggest advantage of evidential reasoning is its ability to explicitly represent uncertainty in the form of a disjunction like *"event A or event B happened with a belief of m"*, without the necessity to assign probabilities to the individual events. Classical probability can not express uncertainty, because assigning the same probability to both events has a slightly different meaning. The relevant notions of TBM are given below, while a complete overview can be found in [13].

Frame of Discernment. The *frame of discernment* (FOD) Ω is a finite set of elementary propositions (or hypotheses) on which beliefs can be constructed. It is also called universe of discourse or domain of reference. Let 2^Ω be its power set. For example, if $\Omega = \{a, b, c\}$ then beliefs can be constructed on all subsets of Ω given as $2^\Omega = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$.

Basic Belief Assignment and Total Belief. A *basic belief assignment* (bba) is a function m: $2^\Omega \rightarrow [0, 1]$ such that $\sum_{A:A\subseteq\Omega} m(A) = 1$. $m(A)$ is also called the *basic belief mass* and represents the belief that the proposition A (a subset of Ω) is true, without supporting any specific subset of A. One can say, it is the belief that event A happened. Because any bba to a subset of A also supports the event being in set A, the *total belief* one can assign to A is given as a function bel: $2^\Omega \rightarrow [0, 1]$ with:

$$bel(A) = \sum_{\substack{B:B\subseteq A \\ B\neq\emptyset}} m(B) \tag{1}$$

$m(\emptyset)$ is not included in $bel(A)$ because it does not explicitly support A, but also supports $\bar{A}$. The total belief can be interpreted as the minimal support for A. The maximal support for a proposition is called *plausibility* and is defined as:

$$pl(A) = \sum_{B:B\cap A\neq\emptyset} m(B) \tag{2}$$

Note the difference between belief bel and plausibility pl, where bel needs B to be a subset of A, while pl only requires B to share some elements with A. It is obvious, that in any case $bel(A) \leq pl(A)$.

It can be seen, that in case of only having propositions A that contain a single element (out of Ω), then the TBM reduces to the standard Bayesian probability distribution. This case can be defined as $m(A) = 0 \; \forall A \subseteq \Omega, |A| > 1$.

Combination of Evidence. Given two bba's m_1 and m_2 from two distinct pieces of evidence, the combined bba of a proposition A can be computed as:

$$m(A) = \sum_{X \cap Y = A} m_1(X) \cdot m_2(Y) \tag{3}$$

In original DS-theory as opposed to TBM, this result was normalized by a factor $1/(1 - m(\emptyset))$, where $m(\emptyset)$ represents the amount of conflict (or contradiction) between the two pieces of evidence defined as:

$$m(\emptyset) = \sum_{X \cap Y = \emptyset} m_1(X) \cdot m_2(Y) \tag{4}$$

This normalization stems from the closed-world assumption made in original DS-theory, that postulates that the actual event that happened has to lie within the FOD and could not be the empty set. As Smets points out in [12], it is much more plausible to adopt an open-world assumption in TBM and accept the fact, that the actual event could be an unknown event. This open-world assumption also much better suits the domain of HCI, as when dealing with humans it is not unusual that they do something unanticipated.

The following example (adapted from [12]) should elucidate the rule of combination and the meaning of conflicting evidences. Suppose a Mr. White has been murdered and we have three suspects: Henry, Tom, and Sarah. Thus our FOD is $\Omega = \{Henry, Tom, Sarah\}$. Imagine we have two witnesses expressing their beliefs about who might be the murderer. Witness 1 states "Henry and Tom could both be the murderer. Sarah is less likely to be the murderer." Whereas witness 2 states "Tom is the one with the strongest motive for murder. Sarah's motive is weaker, but Henry definitely can not be the murderer." The degrees of belief of each witness and the result of evidence combination is given in Table 1. After combining the two evidences, the strongest belief in being the murderer lies on Tom (0.64), followed by the belief (resulting from conflict) that none of the three suspects did it (0.32), followed by Sarah (0.04) and Henry (0.0).

Table 1. Combination of evidences from two witnesses of a murder

		Witness 2		
Witness 1		Henry (0.0)	Tom (0.8)	Sarah (0.2)
Henry, Tom	(0.8)	Henry (0.0)	Tom (0.64)	Ø (0.16)
Sarah	(0.2)	Ø (0.0)	Ø (0.16)	Sarah (0.04)

Decision Making. While beliefs are useful to combine different sources of evidence as described above, they are not directly suited to derive decisions. A decision is better justified by probabilities for each elementary proposition out of the FOD. For this reason the so called *pignistic transformation* (from latin 'pignus': a bet) is usually applied. It transforms a belief function into a probability function over Ω, denoted by $BetP$ and is given by:

$$BetP(\omega) = \sum_{A:\omega \in A \subseteq \Omega} \frac{m(A)}{|A|} \qquad \forall \omega \in \Omega \tag{5}$$

Where $|A|$ denotes the number of elementary propositions in set A. Simply stated, the pignistic transformation distributes the mass of a set A on all its single elements. The probability $BetP(\omega)$ is delimited by $bel(\omega)$ and $pl(\omega)$ such that $bel(\omega) \leq BetP(\omega) \leq pl(\omega)$. So the amount of uncertainty in a hypothesis is just the amount between *bel* and *pl*, often represented as a so-called *belief interval*.

3.2 Fusing Multimodal Inputs with TBM

The traditional evidence theory described above can be used to combine evidences from different sensors to make a decision on individual events described in the FOD. Using the rule of combination from Eqs. (3) and (4) reinforcement, disambiguation, and detection of conflicting evidences are possible. What is missing is a real fusion of inputs coming from different sensors to form an extended concept. For example, if the user wants to select an object o using a verbal deictic reference like "this one" and a pointing gesture, this would lead to beliefs about $a =$ 'select' and $b =$ 'object o'. The fusion system should be able to produce a belief about the combined concept $ab =$ 'select object o'. However the traditional rule of combination results in belief values for the individual concepts a and b, but not taken together. As decribed in [10], the TBM theory can be modified to allow a real fusion of multimodal inputs.

Modification of the Rule of Combination. Key to the modification of TBM is the introduction of tuples as a representation of a combined concept. In the above example instead of just having two events a ('select') and b ('object o'), the tuple (a, b) is used to denote the combined concept. It is important to note the difference between the interpretation of the set $\{a, b\}$ and the tuple (a, b). While $m(\{a, b\})$ is the basic belief that the event was either a or b, $m(\{(a, b)\})$ or short $m(\{ab\})$ is the belief in the combined event ab. Mathematically speaking tuples are the result of a set multiplication operation $\{a\} \times \{b\} = \{(a, b)\}$. Using this cartesian product, the rule of combination from Eq. (3) can now be rewritten [10]:

$$m(C) = \sum_{C:C=(A\times B)\cap\Omega} m_1(A) \cdot m_2(B) \tag{6}$$

Where $(A \times B) \cap \Omega$ allows only those combinations, out of the cartesian product $A \times B$, that are defined in the FOD. One benefit of evidential reasoning is its easy

scalability. Extended to an arbitrary number of sources of evidence the equation becomes:

$$m(C) = \sum_{C=(E_1 \times E_2 \times \cdots E_n) \cap \Omega} m_1(E_1) m_2(E_2) \cdots m_n(E_n) \tag{7}$$

Similarly, the computation of conflict from Eq. (4) becomes:

$$m(\emptyset) = \sum_{(E_1 \times E_2 \times \cdots E_n) \cap \Omega = \emptyset} m_1(E_1) m_2(E_2) \cdots m_n(E_n) \tag{8}$$

Using the strict mathematical definition of a cartesian product, this new combination rule would only produce beliefs over combined concepts (tuples) and not elementary events anymore. To avoid this, Reddy and Basir introduce a neutral evidence '*' that must be part of every sensor's input. So the belief distribution of a sensor must in any case at least contain a belief for '*', meaning that nothing has been detected. Additionally, the combination of evidences obeys the following rules:

- Combination of an evidence with the neutral evidence '*' results in that evidence itself. So $\{a\} \times \{*\}$ results in $\{a\}$.
- The combination of an evidence with itself results in that evidence itself. So $\{a\} \times \{a\}$ results in $\{a\}$.
- The order of evidences in a combined concept is not stressed. That is, $\{ab\}$ is the same as $\{ba\}$.

To elucidate the outcome of this rules, a short example shall be given. Let the FOD be $\Omega = \{a, b, c, d, e, ae, eb\}$. Assume we have two sensors that return belief distributions over the following subsets of Ω:

$$\text{Sensor 1: } \{a\}, \{a, b\}, \{c\}, \{*\}$$
$$\text{Sensor 2: } \{a\}, \{d, a\}, \{a, e\}, \{*\}$$

The combination of these sensors' outputs using the above discussed rules then results in the following table:

Table 2. Example of the adapted rule of combination of TBM based on the cartesian product of sets

	Sensor 2			
Sensor 1	a	d,a	a,e	*
a	a	a	a, ae	a
a,b	a	a	a, ea, eb	a,b
c	Ø	Ø	Ø	c
*	a	d,a	a, e	*

As can be seen in the last row and column, the neutral element preserves beliefs over the original single events. The combinations in the other cells are the result of intersecting the cartesian products with the FOD.

Conceptual Graphs as Representation of Possible Interactions. For an intuitive representation of the FOD, Reddy and Basir propose the use of simplified *conceptual gaphs* (CG) as representation of possible interactions. The CG can be viewed as some form of predefined knowledge, stating what elementary events exist and which of these can be combined. Let the CG be defined as a directed graph (V, E), where V is the set of vertices in the graph (representing the elementary events), and $E \subseteq V \times V$ is the set of edges linking the events. The FOD then simply contains all vertices in the graph and all edges as tuples:

- $A \in \Omega, \quad \forall A \in V$
- $((A \times V \cup V \times A) \cap E) \in \Omega$

Recalling the above example of $\Omega = \{a, b, c, d, e, ae, eb\}$, the CG this FOD could be created from is depicted in Fig. 1. So CGs can be used to represent possible

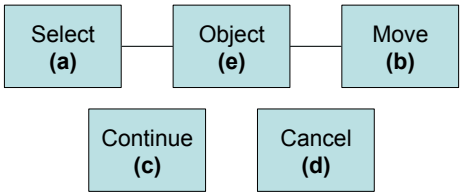


Fig. 1. A simplified conceptual graph relating some basic events

interactions. Together with the modified rule of combination in Eq. (7) and the pignistic transformation from Eq. (5), the theoretical basis for a fusion system of multimodal interactions is laid. However, while Reddy and Basir remain at the theoretical level and in the domain of human robot interaction, our goal is to incorporate this fusion technique into GUI-based applications. For this purpose, the fusion technique must be embedded into a complete system architecture that is able to construct a FOD from a real application, provide the fusion system with actual inputs from sensors, and pass fusion results on to the application. In the remainder of the paper we will describe our current approach.

4 The Proposed Approach

When applying the fusion technique described above in an actual system in order to provide a GUI-based application with inputs, several questions are implied: How do the application and the fusion system act in concert? How are Sensors connected to the system? Is there a way to provide the application with parameters, instead of only raising trigger events? How do we handle different temporal alignments of inputs?

This section illustrates our answers to these questions. First, the application scenario is briefly described before an overview of the whole system and

its architecture is given. After that, processing details exemplified through the application scenario are provided in the subsequent sections.

4.1 Application Scenario

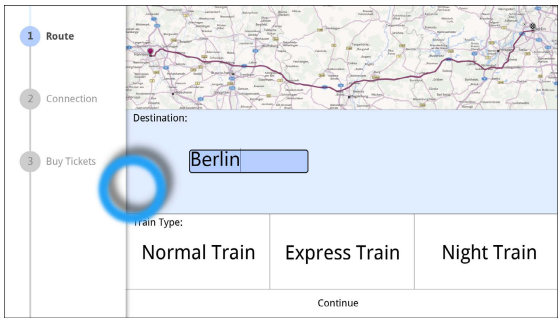
The application that serves as a generalizable example of GUI-based applications is a prototypical train ticket booking system presented on a wall-mounted display. Fig. 2(a) shows a screenshot of the application. To allow for multimodal interaction, the system is equipped with sensors to recognize gestures and speech. An additional location sensor is used to detect a user's approach to the system, to allow for implicit reactions like greeting and taking leave of the user. While the application was designed to be used with conventional ways of interaction (touch, mouse, and keyboard), we show that our fusion system is capable of enhancing it with natural multimodal interaction possibilities. E.g. the user is able to simultaneously use gestures and natural language to select options and to fill in necessary data, while the applied fusion technique combines, disambiguates, and reinforces inputs.

4.2 System Overview

The relevant components of the system and their connections are illustrated in Fig. 3. Recalling the different fusion levels mentioned in Section 2, the fusion system performs a hybrid level fusion, as connected sensors do not provide final decisions but belief distributions. The overall fusion processing is twofold. First, the application sends a state description to the fusion system (① to ⑪ in Fig. 3) that is used to dynamically reconfigure the fusion process for the current possible interactions. Second, there is the input fusion pipeline (① to ④) that receives recognition events from the connected input sensors and passes them through the actual fusion process. Finally, input events are sent to the application when indicated by the fusion result. Note that the number of sensors is not restricted to the one shown in Fig. 3 as the fusion system allows an arbitrary number of sensors. From a technical perspective, the major parts of the system (i.e. sensors, fusion system, and application) are loosely coupled via a message based middleware and all messages are sent in XML format, to ensure flexibility and exchangeability. The following sections describe the two parts of processing, namely the connection to the application and the input processing pipeline in detail (cf. Fig. 3).

4.3 Connection to the Application

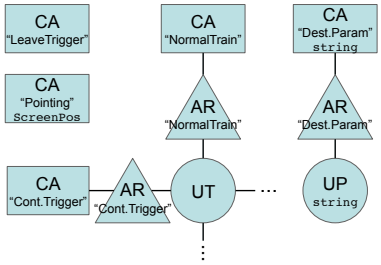
In order to realize a multimodal input fusion for the application, there must exist some kind of interface between the application and the fusion system. Whether it is a standalone application or a fully adaptive dialog management system, all that is needed is a description of all possible interactions that are allowed in the current state. The description itself is independent of the modalities used for input, as the mapping between the interactions expressed by the user and these interactions is the principal task of the fusion system.



(a)

```
<State>
  <Action ID="LeaveTrigger"/>
  <Action ID="Pointing">
    <Parameter>
      <ScreenPosition/>
    </Parameter>
  </Action>
  <Action ID="DestinationParam">
    <Parameter>
      <String/>
    </Parameter>
    <Representation>
      <ROI x="420" xDim="1500" y="360" yDim="360"/>
      <String value="Destination"/>
    </Representation>
  </Action>
  <Action ID="NormalTrain">
    <Representation>
      <ROI x="420" xDim="500" y="720" yDim="260"/>
      <String value="normal"/>
    </Representation>
  </Action>
  <!-- other train types -->
  <Action ID="ContinueTrigger">
    <Representation>
      <ROI x="420" xDim="1500" y="980" yDim="100"/>
      <String value="Continue"/>
    </Representation>
  </Action>
</State>
```

(b)



(c)

Fig. 2. (a) Screenshot of the prototypical train ticket booking system allowing to specify a destination and select a train type. The blue circle visualizes a pointing gesture. (b) The screen's abstract interaction description (AID) in XML. (c) The CG created from the description. Each edge linking two nodes implies a possible combination of events.

Abstract Interaction Description. The interaction description provided by the application should be able describe the basic actions that are common practice in normal applications, including the triggering of actions and the providing of parameters. For this purpose we use an *abstract interaction description* (AID) in XML format. Our current implementation allows for XML elements to describe triggers and typed parameter actions. Additionally, the representation of an action on screen can be described in terms of the *region of interest* (ROI) and its textual representation. Fig. 2(b) shows an example from the application scenario.

The upper two `<Action>` elements describe triggers that do not provide a `<Representation>` description and therefore exhibit *implicit interactions*. The other actions exhibit a ROI and a string representation, implying *explicit interactions*. This means, they are actually part of the application's graphical

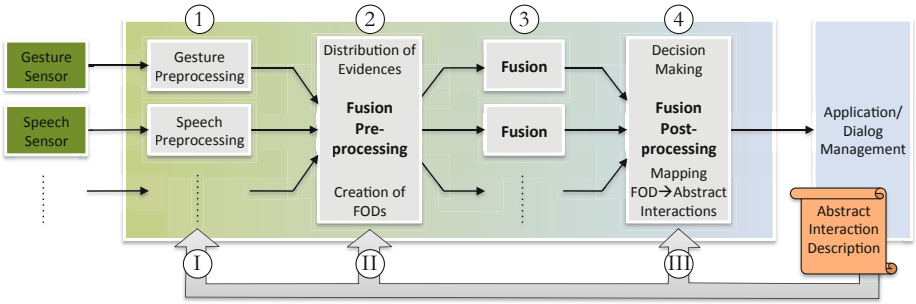


Fig. 3. The main components of the fusion system. Configured by a state description of the application, multiple parallel fusions based on evidential reasoning are performed. Based on the fused results a final decision is made and interaction events are passed over to the application.

presentation, e.g. a button on screen, that can directly be referred by the user. The 'DestinationParam' element describes an action with a `string` parameter used to give the travel destination.

The AID not only describes the current possible interactions of an application towards the fusion. It also serves as description of events that can be raised by the fusion towards the application. So it constitutes the complete interface between the fusion system and the application at any instant. The `<Parameter/>` elements deserve special attention. As long as they are part of the AID sent by the application, they simply state the type of parameters expected by the application. E.g. the action 'DestinationParam' in Fig. 2(b) has a parameter of type `string` without any value. Then, during the fusion process they get filled with actual values, before they are finally returned to the application in an interaction event.

As mentioned above, the AID can not only contain *explicit interactions*, that the user should be able to perform willingly. It can also contain *implicit interactions* that are usually not meant to be intentionally executed by the user, e.g. the 'LeaveTrigger' in Fig. 2(b). These kind of interactions can often be found in ubiquitous systems (e.g. smartphones that react upon being raised to the ear) but also more common features of applications like giving feedback on a pointing action are a special case of implicit interactions. But how can these interactions be invoked? Within our approach, this is realized by an additional mapping description provided by the application. This description simply states which actions (referred by their IDs from the AID) are triggered by which sensor input events. In the above example, there is a mapping defined for the 'LeaveTrigger' action to be invoked whenever the location sensor detects the leaving of the user.

Fusion System Configuration via an AID. Whenever the application sends an AID the fusion system dynamically reconfigures to account for the change in possible interactions. As depicted in Fig. 3, there are three stages, where the

AID is used: the sensor preprocessings Ⓘ, the fusion preprocessing Ⓘⓘ, and the fusion postprocessing Ⓘⓘⓘ.

Configuration of Sensor Preprocessings Ⓘ For each sensor that is connected to the system, there exists a dedicated preprocessing component. Herein, the AID is used to restrict the sensor messages thrown into the fusion process to those relevant in the current application state.

Configuration of Fusion Preprocessing Ⓘⓘ Here, the AID is transformed into a conceptual graph from which FODs are created. These control the evidential reasoning performed in the actual fusion process(es). Details on this transformation are given in the following section.

Configuration of Fusion Postprocessing Ⓘⓘⓘ During the creation of the conceptual graph in the preprocessing, mappings from the CG back to the original AID are stored in the postprocessing component. This mappings are used later to create actual application events from the fusion results.

From the Interaction Description to FODs. One of the most important aspects of evidential reasoning is the creation of elementary propositions as defined by the frame of discernment (see Section 3). We decided to use a *common meaning representation* (CMR) as operational base of the fusion system that makes the fusion itself independent of actual used sensors and modalities. Thus, the AID coming from the application is first transformed into a conceptual graph stating what elementary events exist and which of these can be combined by the fusion (see Section 3.2). The following elements currently defined in the CMR are sufficient to express every multimodal input intended to evoke any of the AID-actions in an abstract manner:

Action Reference (AR). A reference to an action that is identified via its ID.

Undetermined Trigger (UT). A universal trigger that is not (yet) associated with a specific action.

Undetermined Parameter (UP). A typed parameter that is not (yet) associated with a specific action.

Complete Action (CA). A complete action (including eventual parameters) that is triggered.

The transformation itself is currently done using rules, that define for all possible types of actions, which elements are created in the CG and how they are linked. E.g. as depicted in Fig. 2(c) the 'DestinationParam' action from above is transformed into an AR of the action linked with both, an UP of type `string` and a CA with a `string` parameter. The created CG represents the overall FOD upon which the actual fusion is performed.

The Necessity for Multiple Parallel Fusions. The depicted CG in Fig. 2(c) is not a connected graph but contains several disjoint subgraphs. This differs from the CGs given in [10], as they only consider connected graphs. But what is the

implication of that? It means, that we do not only consider solely combinable events, but also events that have a meaning on their own (e.g. the 'LeaveTrigger' action) without the need to be part of a combination. Additionally, the intuitive impact of having multiple disjoint subgraphs is, that they represent events completely independent of one another. From this it follows that fusion can not be performed on a single FOD but must be performed with multiple FODs in parallel. Otherwise concurrent occurrences of independent events would often lead to the maximum belief being assigned to the conflict $m(\emptyset)$ when applying Eq. (8). For this reason, in our approach, the constructed CG is partitioned into its disjoint subgraphs and a separate FOD and fusion component is created for each of them.

4.4 Input Processing Pipeline

Now that the system is configured for the current state of the application, it is finally ready to receive and fuse inputs. The following sections elucidate the different steps of the input processing pipeline as indicated by arabic numerals ① to ④ in Fig 3.

Connection to Sensors. Currently we use an off the shelf speech recognition component, and gesture and location detection sensors from consortium partners. All sensors communicate with the system via XML messages. These messages usually contain recognition results together with beliefs or confidence scores. The different sensor preprocessing components ①, that have already been configured for the current AID, create events over the CMR (see Section 4.3) and assign beliefs. So they are 'translating' modality specific recognitions into elementary events of the FODs. E.g. the speech preprocessing creates an undetermined trigger event, when the user says "this one". When there are actions with a representation description, action references can be created according to the ROI (pointing gesture) or the specified string representation (speech). Additionally, parameters are filled when they match the specified type, e.g. a non-keyword string recognized by the speech sensor gets assigned to an undetermined parameter event of type `string`.

Fusion Preprocessing. Whenever a sensor preprocessing signals a new sensor output, the fusion preprocessing ② redistributes the incoming belief distributions among the currently available fusion components. I.e. whenever a signaled belief of a sensor matches an element of a fusion's FOD, the belief is added to the input of that fusion. In order to preserve valid belief distributions that satisfy $\sum_{A:A\subseteq\Omega} m(A) = 1$ (basic beliefs must sum up to 1), every dropped out belief mass gets assigned to the neutral belief '*'.

The Actual Fusion. Using the constructed belief distributions over their FODs, the fusion components ③ perform the combination of evidences using Eqs. (7) and (8) to compute combined beliefs and the overall conflict. For visualization

and debugging purposes, total beliefs and plausibilities (*bel* and *pl* from Eqs. (1) and (2)) are also computed. Finally the pignistic transformation $BetP$ from Eq. (5) is applied, before the result is handed over to the fusion postprocessing ④.

As an example form the application scenario imagine a situation where the user simultaneously points towards the screen and makes the utterance "I want to take a normal train". Given the situation that the surroundings are quite noisy, the speech recognition only assigns a belief of 0.4 to the event being the action 'NormalTrain'. The gesture recognition may have some ambiguity about wether the pointing refers to the normal train or the express train due to limited resolution capabilities of the vision based system. Therefore it assigns a belief of 0.8 to the event being both, a reference to 'NormalTrain' and 'ExpressTrain' at the same time. Using the evidential reasoning approach, this ambiguity can be resolved via combination and reinforcement of the given evidences. Fig. 4 shows the actual visualization of the fusion system in this situation. The highest pignistic value (0.32) is correctly assigned to the combined event of the action and reference to 'NormalTrain', resulting in a selection of the normal train in the application.

Fig. 4. Screenshot of the fusion system, where low confidence and ambiguous inputs (top left) are combined resulting in the highest probability for the correct event (lower right)

Temporal Fading of Inputs. May it that the user performs inputs in a sequential manner (using *integration patterns* [8]) or that sensor processing's take some while, it is quite rarely (if ever) that different sensors raise recognition events simultaneously. Since the evidential reasoning is ad hoc, there is a need to extend the temporal scope of such events. Thus our approach uses *temporal fading*, similar to Pfleger's way of using activation values [9]. After new evidences are assigned their initial belief values, they continuously get decreased over time,

while at the same time the belief in the neutral evidence '*' is increased. When all the beliefs of a sensor output have reached 0 (and $m(*) = 1$), the output is finally removed from the fusion system. In other words, the fusion system has a form of memory that fades over time. Currently the time interval for all evidences is fixed at 1.2s, resulting from initial tests with the system.

Fusion Postprocessing. When a fusion component completed its pignistic transformation, the resulting list of probabilities is handed over to the fusion postprocessing (cf. Fig 3 ④). Here, the event with the highest probability is mapped back from the FOD to the actions currently defined in the AID. If the event carries parameters, they get filled into the XML representation of the action, before it gets sent to the application. There are currently three cases, where no action is sent to the application: either the 'winning' event has no mapping defined (e.g. a single reference, without an accompanying trigger), or the neutral or conflicting evidence is the one with the highest probability. As a potential future enhancement, conflicting evidences could be communicated to the user, to make him become aware of the systems inability to decide on a particular action.

5 Summary and Outlook

As depicted above, our applied fusion system is able transfer the capabilities of evidential reasoning for multimodal interaction into the generic domain of GUI-based applications. Using a simple yet powerful common meaning representation of interactions and multiple parallel fusions, the system is able to perform some of the most common tasks of GUI-based applications (e.g. triggering actions, providing parameters) in a natural, and where applicable, multimodal way. At the same time the robustness of input recognition is increased by the combination, disambiguation, reinforcement, and conflict detection capabilities of the adapted TBM; capabilities rarely found in other rule based approaches presented in Section 2. As already shown in [10], using the adapted TBM is more robust at representing and combining events as compared to purely probability based approaches and even traditional DS-theory. Admittedly the benefit of the approach is highly dependent on the capabilities of the used sensors and their preprocessings to produce sound belief distributions.

Initial tests with the implemented system are promising, though a real evaluation needs further work. It remains unclear, if the developed abstractions are sufficient for most use cases, as only further testing with real users can show. Aside from that, future work will explore the possibilities to adapt to different types of users, as there is evidence that users are highly different in their way of interacting multimodally [8]. An obvious yet simple approach, currently in progress, would be to adapt the temporal fading of events. While the presented approach per se is not able to detect complex command sequences, it could be combined with one of the existing rule based systems, where the sequence of produced events from the presented approach serves as input. Such a multi-layer

system could be the ideal solution for input fusion in companion systems, combining robustness and reliability with a sophisticated understanding of humans' multimodal behavior.

Acknowledgements. This work is originated in the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Atrey, P., Hossain, M., El Saddik, A., Kankanhalli, M.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems 16, 345–379 (2010)
2. Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J.: Quickset: multimodal interaction for distributed applications. In: Proc. of Multimedia 1997, pp. 31–40. ACM (1997)
3. Dumas, B., Ingold, R., Lalanne, D.: Benchmarking fusion engines of multimodal interactive systems. In: ICMI-MLMI 2009: Proc. of the 2009 International Conference on Multimodal Interfaces, pp. 169–176. ACM (2009)
4. Dumas, B., Lalanne, D., Ingold, R.: Description languages for multimodal interaction: a set of guidelines and its illustration with smuiml. Journal on Multimodal User Interfaces 3, 237–247 (2010)
5. Dumas, B., Lalanne, D., Oviatt, S.: Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In: Lalanne, D., Kohlas, J. (eds.) Human Machine Interaction. LNCS, vol. 5440, pp. 3–26. Springer, Heidelberg (2009)
6. Holzapfel, H., Nickel, K., Stiefelhagen, R.: Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. In: Proc. of ICMI 2004, pp. 175–182. ACM (2004)
7. Nigay, L., Coutaz, J.: A generic platform for addressing the multimodal challenge. In: CHI 1995: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp. 98–105. ACM (1995)
8. Oviatt, S.: Multimodal Interfaces, ch. 21, 2nd edn., pp. 413–432. CRC Press (September 2007)
9. Pfleger, N.: Context based multimodal fusion. In: Proc. of ICMI 2004, pp. 265–272. ACM (2004)
10. Reddy, B.S., Basir, O.A.: Concept-based evidential reasoning for multimodal fusion in human-computer interaction. Appl. Soft Comput. 10(2), 567–577 (2010)
11. Sharma, R., Pavlovic, V., Huang, T.: Toward multimodal human-computer interface. Proc. of the IEEE 86(5), 853–869 (1998)
12. Smets, P.: The combination of evidence in the transferable belief model. IEEE Trans. Pattern Anal. Mach. Intell. 12(5), 447–458 (1990)
13. Smets, P.: Data fusion in the transferable belief model. In: Proc. of the Third International Conference on Information Fusion, FUSION 2000, pp. PS21–PS33 (2000)
14. Wendemuth, A., Biundo, S.: A Companion Technology for Cognitive Technical Systems. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) COST 2102. LNCS, vol. 7403, pp. 89–103. Springer, Heidelberg (in press or to appear 2012)

Fusion of Fragmentary Classifier Decisions for Affective State Recognition

Gerald Krell[1], Michael Glodek[2], Axel Panning[1],
Ingo Siegert[1], Bernd Michaelis[1],
Andreas Wendemuth[1], and Friedhelm Schwenker[2,⋆]

[1] Institute for Electronics, Signal Processing and Communications,
Otto-von-Guericke University Magdeburg, Germany
{firstname.lastname}@ovgu.de
http://www.iesk.ovgu.de
[2] Institute of Neural Information Processing, Ulm University, Germany
{firstname.lastname}@uni-ulm.de
http://www.informatik.uni-ulm.de

Abstract. Real human-computer interaction systems based on different modalities face the problem that not all information channels are always available at regular time steps. Nevertheless an estimation of the current user state is required at anytime to enable the system to interact instantaneously based on the available modalities. A novel approach to decision fusion of fragmentary classifications is therefore proposed and empirically evaluated for audio and video signals of a corpus of non-acted user behavior. It is shown that visual and prosodic analysis successfully complement each other leading to an outstanding performance of the fusion architecture.

Keywords: Emotion recognition, human computer interaction, multimodal data.

1 Introduction

Human-computer interaction (HCI) is increasingly based on user inputs detected in speech, facial expressions or gestures from audio and video [24], not to mention physiological signals [30,31]. These kinds of user input can directly control the interface of a computer application. Another, more abstract problem considered here is to estimate the user's *affective state* from the user input. This could be useful to control the communication strategy of the interacting computer. The application could, for example, change to a supporting mode when the user operating the system is found in a desperate situation [29]. Other well-known

⋆ Jörg Frommer (Department of Psychosomatic Medicine and Psychotherapy, Magdeburg) and Dietmar Rösner (Department of Knowledge and Language Engineering, Magdeburg) with their teams deserve our special thanks for providing the LAST MINUTE corpus [20]. Gerald Krell and Michael Glodek contributed equally.

F. Schwenker, S. Scherer, and L.-P. Morency (Eds.): MPRSS 2012, LNAI 7742, pp. 116–130, 2013.

applications are detection of drowsiness in driver assistance systems [27,1] or pain states in medicine [15].

Reliable estimation of the user state in HCI is still a challenging task in pattern recognition [17]. One particular field of intensive research is emotion recognition. For many years, the focus has been set on acted data, also due to lack of available non-acted datasets [33,17]. In the case of acted emotional data the label (ground truth) is well defined by instructing the actor and therefore, the training of supervised classifiers can easily be realized. The acting person behaves usually in an ideal manner, is positioned in front of the camera, speaks distinctly, etc.

However, non-acted "real" data raises the problem of reliable ground truth. Experiments on detection of spontaneous, affective valence or arousal face the problem of rather uncertain training data, in opposite to many other machine learning applications. What is more, the expressiveness level in real situations is usually lower than for acted data [4,25]. Real test persons do not behave ideally from the data acquisition point of view: they may not speak distinctively, turn away from camera or hide parts of their face by hair, glasses or hands.

The alternative of a subjective post-process labeling often turns out in bad inter-rater reliabilities, especially when more than one modality is involved. Currently the first option of purposed eliciting affects seems more promising and is widely used [24]. One solution is to elicit a test person purposely towards showing any kind of affective behavior. The idea is that a ground truth is defined by the experimental setup and/or applied stimuli as done in [26]. But one has to bear in mind that eliciting the affect cannot guarantee the actual appearance of the targeted emotion.

Most common approaches perform fusion on decision level [10,16,14,28]. In facial expressions, for instance, the extraction is based on feature point positions from which the movement of individual muscles can be derived. This principle is applied in the facial action coding system (FACS) of Ekman [6] or similar approaches as in [16]. Another approach is the analysis of active regions (eye, mouth, forehead) using black-box-like descriptors produced by discrete cosine transform [10].

For analysis of prosody in speech, Mel-frequency cepstral coefficients (MFCC), pitch and energy/intensity are the most common descriptors [2]. They are used to generate prosodic features in order to derive data describing the affective user state.

Gesture recognition is already commonly used to directly control application interfaces. Affective state recognition from hand and body gestures is an active field of research [13,22].

In summary, it becomes clear that recognition of the affect state is still missing standardized analysis due to the fact that scientific findings are still in its infancies. In the following, the proposed classifiers for visual and prosodic features with their specific properties and the fusion method are introduced. In the experimental part, the applied corpus and classification results for two selected key events are presented.

2 Fusion of Fragmentary Modalities

In real HCI systems, the classification in different modalities is quite vague. What is more, the information in the channels might not be continuously available, e.g. because the face of a person is occluded (facial expression channel) or there are no utterances (prosodic channel) at the considered time step. Given a set of n possible user states $E = \{s_1 \dots s_i \dots s_n\}$ and assuming the user to be in a certain user state $S \in E$ it should therefore be investigated how information of such channels can be fused to increase reliability of the recognized user state $\hat{S} \in E$ (see Fig.1).

According to our assumption the actual (real) state S cannot directly be measured, but we know the event a-priori from the experimental set-up as ground truth. Only the observable features detectable in the data modes (audio, video, physiological signals etc.) are available from outside. One problem considered in this paper is that it is very typical for these features to occur more or less occasionally, such as

- prosodic features are only available when the user speaks,
- gestures are only detected when typical hand movements occur,
- facial expression are usually temporary.

Distinctiveness of features strongly depends on the temperament, gender, cultural background etc. of the particular test person. Another challenge is that real sensors do not always produce useful data, such as in case

- user heads off from camera,
- face is hidden by hands,
- mouth speaking movement overlays facial expression.

The missing information in the channels is indicated as gaps in Fig. 1. Observable features are to be retrieved from different modalities such as audio, video or physiological signals by suitable classifiers. Assuming that these data modes are taken time-synchronously, but are not always available in every considered time step, we first perform a classification based on each modality by a suitable method when the signal is present. The output of classifiers is also fragmented and the reliability (confidence) of the classification varies severely.

In the second step, a fusion algorithm based on a Markov fusion network (MFN) estimates a continuous output $\hat{S}$ (regular sampling) based on the fragmented output of the classifiers.

2.1 Audio-Visual Classifiers for Observable Features

The modalities used in this paper providing observable features are the audio and video channels. Audio is analyzed with respect to its prosodic information and the video channel with respect to the facial activities and hand gestures. It is assumed that the proposed framework can be extended to also include additional channels such as physiological signals or body gestures.

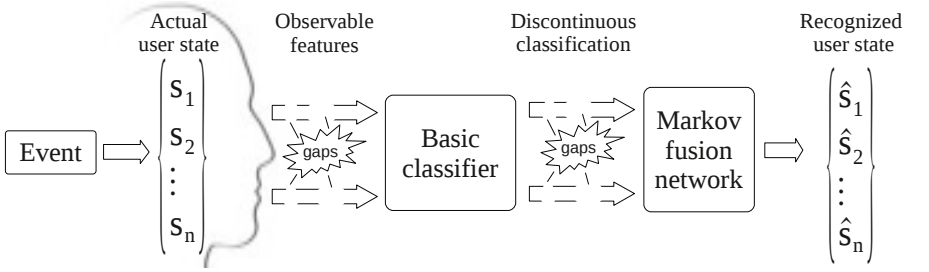


Fig. 1. Key events induce certain affective user states. Basic classifiers rely on observable features of different channels. Fragmentary classifier decisions are then fused by the proposed MFN to estimate user state continuously.

In the following we give a short overview on the observable features and how a basic classification is done, providing the input for proposed fusion algorithm.

Audio Classifier. The actually used prosodic features are generated from commonly used short-term acoustics. Therefore, we generally first separate each utterance spoken by the user and apply a Hamming window to get short-time stable acoustics.

Afterwards, MFCC, their deltas and acceleration using the Hidden Markov toolkit (HTK) [32] are calculated. The MFCCs describe the Mel-scaled short-term power spectrum. Deltas and accelerations are used to include the dynamics of the speech signal. We also include a zero mean static coefficient, which provides an indication for the overall level of the speech frame. These features have been proven to provide good recognition results for speech [9], emotion [23] and speaker recognition [7].

Visual Classifiers. The considered facial activities are mouth deformations, eyebrow actions, eye blink and global head movement. This information is mainly generated by measuring facial distances d_i (see Fig. 2, left) and by measuring the head position with an optical 3-D scanner using a head model. More details on facial expression analysis can be found in [19] and [18].

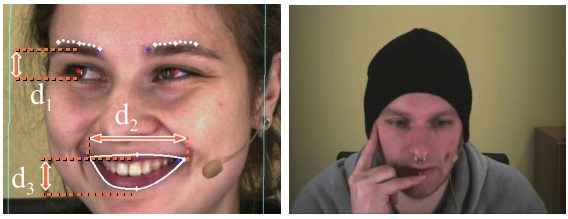


Fig. 2. Left: distance features d_1, d_2 and d_3 used for facial expression analysis (experiment E35); right: self-touch gesture (E59)

The detection of gesture aims at recognizing self-touch actions (i.e. hand touches the face, see Fig. 2, right) which is an important information from a psychological point of view in the context of affective state recognition [12]. Self-touch detection is based on skin color. After extraction of skin-blobs in the image, a connected component analysis determines whether a hand approaches the face and eventually touches it. Self-touch region is also estimated, but actual investigations are restricted on the binary decision whether a self-touch is currently present or not. A detailed description of the method can be found in [21].

On the one hand we assume that the affective state is stable over a certain time and dynamics is limited. On the other hand, we believe that the affective state is reflected by the dynamics of observable features. We therefore consider a time window instead of using single independent frames for the estimation of user state as other approaches do (mostly for acted databases). The time window considers the dynamics of user behavior. The influence of window size is explicitly evaluated in [19].

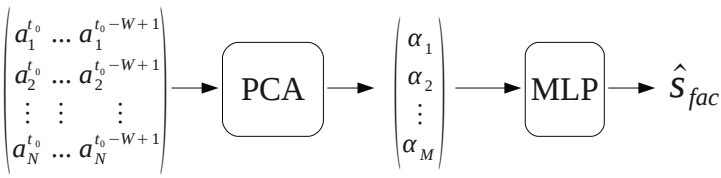


Fig. 3. The input features at time t_0 are transformed by M most important eigenvectors from PCA in feature space providing input of MLP for classifying the facial affective state $\hat{S}_{fac}$

A linear model is established by Principal component analysis (PCA) [3] on feature data, where the input values at time step t are provided by $N \cdot W$ values of observable visual features $a_1^t, a_2^t, \ldots, a_N^t$ in a time window of size W as shown in Fig. 3. A standard Multi-layer perceptron (MLP) is trained to classify the affective state on the basis of M most significant principal components $(\alpha_1, \alpha_2, \ldots, \alpha_M)$. The training of the neural network is considerably simplified by the PCA because the number of inputs is much smaller than the total number of features: $M << N \cdot W$. The output of the MLP is a continuous value in the range $\langle 0, 1 \rangle$ providing a confidence value when applying a threshold to classify a certain affective state $\hat{S}_{fac}$ based on facial activities.

2.2 A Markov Fusion Network

The fusion of decisions from different modalities in non-acted real-world data bears a couple of challenges. Not only the channels are, in general, recorded with different frame rates, but the fusion has as well to deal with the event of

missing decisions at a certain time step, due to sensor failure or the inherent characteristics of the modalities. Furthermore, when leaping from the laboratory to the real world, the uncertainty and missing pre-segmentation of the classification comes into focus of fusion.

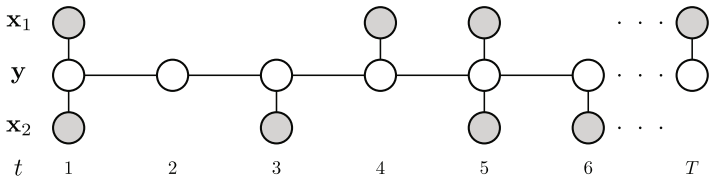


Fig. 4. Graphical model of the MFN. The estimates y_t are influenced by the available decisions x_{tm} of the source m and $t \in \mathcal{L}_m$ and the adjacent estimates y_{t-1} and y_{t+1}.

As a result, the fusion has to deal with an irregular temporal sampling of decisions resulting because different sources have their individual reliability and temporal validity. In order to cope with these aspects we make use of a Markov fusion network (MFN)[8], which is designed to combine decisions from multiple sources with temporal dependencies. The model is originated from the application of Markov random fields in image processing [3,5].

According to the model, the value y_t is defined as the estimated decision obtained by combining the streams of different sources. The streams are given by $\mathbf{x}_m$ where $m \in M$ is the index of a source. Fig. 4 shows the corresponding graphical model on the example of two sources $\mathbf{x}_1$ and $\mathbf{x}_2$. According to the graphical model the estimates y_t are connected in a chain representing the temporal dependency of the single decisions. If a classifier of a stream provides a decision x_{tm}, a link to the final estimate y_t is added to the graph.

The probability of the MFN is defined by two potentials Ψ and Φ. The potential Ψ is obtained by accumulating the M modalities Ψ_m and enforces the estimate y_t to be equal to the decisions x_{tm} and is defined by

$$\Psi = \sum_{m=1}^{M} \Psi_m = \sum_{m}^{M} \sum_{t \in \mathcal{L}_m} k_{tm}(x_{tm} - y_t)^2, \tag{1}$$

where $\mathcal{L}_m$ is the set of time steps in which a decision for m is available, and, k_{tm} is a parameter defining the strength of its influence. In case a prediction of the reliability of a channel over time is given, k_{tm} may change over time. However, in the current setting we will restrict ourself to the case in which k_{tm} is constant over time. The second potential Φ enforces temporal similarity and is given by

$$\Phi = \sum_{t=1}^{T} \sum_{i \in N(t)} w_{t-i}(y_t - y_{t-(1-2i)})^2 \quad \text{with} \quad N(t) = \begin{cases} \{0\} & \forall t = 1 \\ \{1, 0\} & \forall 1 < t < T \\ \{1\} & \forall t = T \end{cases} \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^{T-1}$ is weighting the cost of a difference between two adjacent nodes. The parameter $\mathbf{w}$ can be set if additional domain knowledge is available at design time, e.g. to weaken similarity in case an extraordinary event is induced by the computer and a change in the estimated decision is likely.

The joint distribution of the estimated vector $\mathbf{y}$ given the decisions of the modalities is defined by

$$p(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_M) = \frac{1}{Z} \exp\big(-\frac{1}{2}(\Psi + \Phi)\big), \tag{3}$$

where Z is the partition function which normalizes the probability such that the integral is equal to one. In order to determine the most probable estimates of $\mathbf{y}$, it is not mandatory to calculate the probability of the model itself. Instead the mode of the likelihood can be obtained by maximizing the log-posterior probability using gradient ascent.

3 Experimental Results

In the past, different approaches have been studied for multi-modal affect recognition. Besides acted multi-modal data, real affective data have been increasingly used [34,14]. Emotional elicitation can be done in different ways. One option is to present images or videos with a contents directly evoking emotions (e.g. [26]). Compared to this approach, we generate a certain affective user state by a rather mundane set-up. The subjects are faced with events typical for real HCI environments, such as unexpected system outputs or stress due to missing interface options.

For the reported investigations, we used data of the LAST MINUTE corpus (LMC) [20], which is described shortly to show how the events in Fig. 1 are generated.

3.1 The Corpus

The LMC contains multi-modal recordings which are taken during a "Wizard of Oz" experiment [11]. The test person interacts with a system appearing as an HCI system, but in fact the application is controlled by a human operator (the so-called wizard) not in view with the subject. The wizard instructs the subject by synthesized speech and visual outputs on a computer screen.

The subjects are briefed that one goal of the experiment would be the test of a new natural language communication interface. Using voice commands the subjects are to prepare a fictional voyage to the unknown place "Waiuku", assemble suitable baggage and clothing. The task contains planning, strategy change and re-planning and is designed to generate affective material for prosody, gesture, facial expressions and linguistic analysis.

The experimental data were recorded by hardware-synchronized cameras and microphones. A detailed description and technical specifications can be found in [20].

We focus on two key events of the experiment ($n = 2$, Fig. 1), where the user is set either in the *Baseline* (BL) or *Challenge* (CH) state. BL is the phase in the experiment after 5-10 minutes when the test person has been adapted to the experimental situation. The CH event happens when the system creates mental stress by suddenly requesting a strategy change (exceeding baggage limit) from the test person who then should be in an aroused mood. This event occurs approximately 15-20 minutes after the beginning of the experiment and brings the test person to re-arrange the assembly of baggage. The stress period is assumed to last approximately two minutes according to the experiment time line. Neither we can be sure about the real stress factor for the particular subject, nor can we assure the real duration of any higher stress level, but the recognition system should be able to classify the key event as reliable as possible.

3.2 Basic Classification

For basic classification (see Fig. 1), dedicated classifiers for visual activities, prosody and gesture have been designed. Fig. 5 shows the temporal availability of features providing the input of these classifiers during the periods of the key events for one typical test subject (E35).

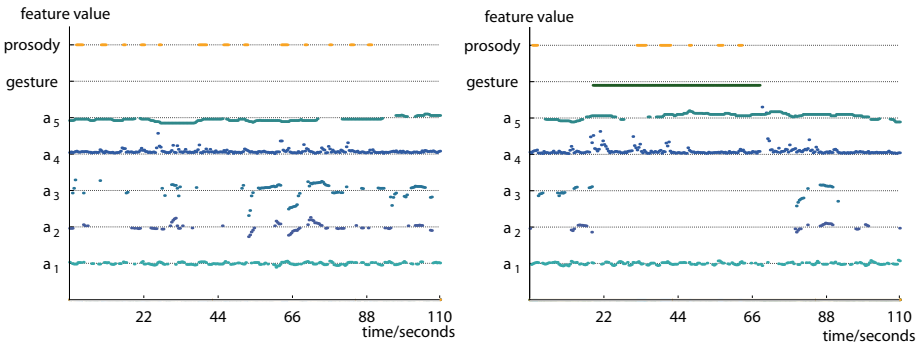


Fig. 5. Observable features for BL (left) and CH (right) of E35. Facial measures $a_1 \cdots a_5$ are given by their normalized deviation from the mean value. Gesture: line indicates self-touch. Prosody: line indicates utterance.

The facial activities (derived from th facial distances d_i, see Fig. 2, left) are mean of left and right brow position (a_1), mouth width (a_2)and mouth height (a_3). The right diagram of Fig. 5 shows that geometrical data of the mouth (a_2, a_3) is not available when this region is hidden by the hand of the user in a self-touch situation. Additionally head position (a_4) and eye blink frequency (a_5) have been estimated.

For prosody, the period in which the user speaks and prosodic features can be derived is indicated.

Gestures are only evaluated regarding the self-touch event. In the example of E35, only in the CH period (diagram on the right) a self-touch occurs.

Prosodic Results. The relation of speaking time for system and user is given in Table 1 to get a feeling about amount of available speech data. Because the system did not allow to barge-in, the overall length of utterances is manageable and thus acoustic cues appear only in a cumulated manner. All wizard utterances have been blanked out.

The speech material comprises 41 speakers, also including subjects, who were not used for visual classification to raise the usable amount of data. On basis of a feature vector with 36 elements for every 15ms and a time window of 25 ms a hidden Markov model (HMM) with 3 states and over-all 81 Gaussian mixture components (full) was trained. The second column of Table 2 shows the recognition rates for the prosodic modality for 13 subjects of the whole corpus where also visual analysis was available. These results where achieved by leave-one-speaker-out (LOSO) validation, where 40 speakers have been used for training and the remaining one for testing and by repeating it ten times for each speaker to obtain a ten-fold validation.

Table 1. Average speaking-time of the wizard and test person for both events with a length of 110 seconds

Event	Wizard	User
Baseline	35.2 ± 6.8s	15.6 ± 4.2s
Challenge	63.3 ± 9.3s	10.5 ± 5.3s

Visual Results. Five facial features (including head pose) according to Fig. 5 in a time window of size $W = 15$ (0.6 s) totally produced 75 features. The feature correlation has been considerably reduced by selecting $M = 4$ principle components as input for the MLP.

The results for the individual subjects are displayed in columns 3 and 4 of Table 2 when applying a threshold of 0.5 to the neuronal outputs for class selection. The results are given separately for BL and CH showing that BL can be detected with higher reliability than CH. With an average recognition rate of 65.7 % for BL and 44.9 % for CH (totally 55.27 %) it becomes obvious that a final decision on affective user state can be hardly made on this single modality.

The low recognition rate for some test persons is typical because some people do not express their emotions at all and also nose glasses or facial hair let the facial recognition fail.

3.3 Decision Fusion

The experimental results of the proposed multiple classifier are based on the basic classifier decisions described in the previous section related to the recordings of 13 participants of the LMC for 110 seconds of the key events BL and CH.

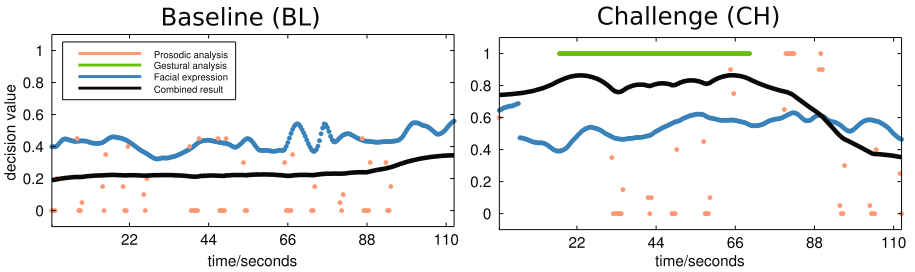


Fig. 6. Representative output of the MFN in the example of E35. The BL event is shown on the left (target class is zero) and the event CH on the right (target class is one). The decisions provided by the facial expression classifier (blue), the prosodic classifier (orange) and the gestural analysis (green) are utilized to estimate the final decision (black) using the MFN. The estimate mediates between the available decisions with additionally taking their temporal distances into account.

Figure 6 shows the output generated by the MFN together with the decisions of the basic classifiers for a typical test person. The input decisions, provided by the prosody, gestures and the facial activities, are depicted as orange, green and blue dots (in monochrome print-out the dots appear in increasingly darker shades of gray). Each of the modalities shows a characteristic temporal pattern.

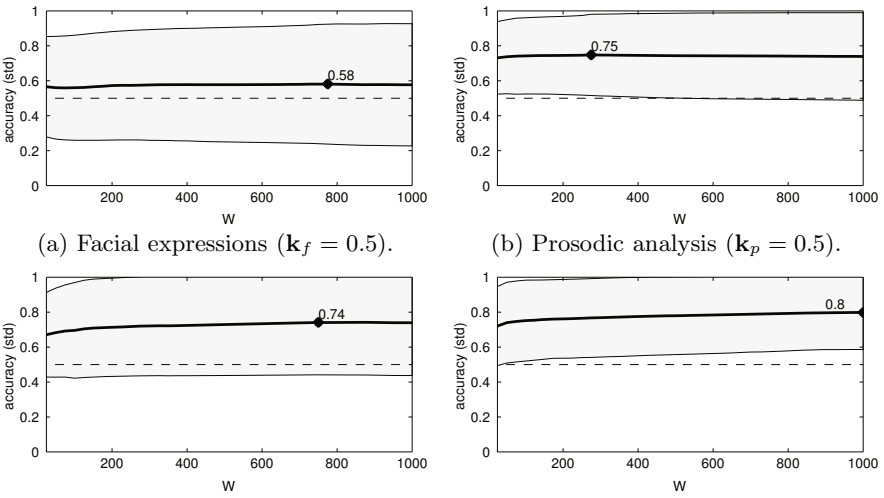
The prosodic decisions are fragmentary because results of the audio channel are only available in case of a verbal utterance of the test person.

The gesture-based decisions provide evidence only for the CH class. It is taken into account for the time interval in which the gesture (i.e. the touching of the own face) is recognized. A corresponding event can be seen on right-hand side of figure.

The facial activities, also derived from the video channel, are almost continuously available because the subject is recorded from a frontal perspective. The combined result is depicted by a solid black curve. It provides a non-fragmented decision for all time-step, and is based on influences from all modalities.

The first experiment addresses the MFN performance on single modalities. Fig. 7a and 7b show the frame-wise accuracy averaged over the subjects (thick black line) evaluated based on the re-estimated decisions of the classifiers using the facial expressions and prosodic analysis (the class prior of 0.5 is indicated by dashed horizontal line).

While the parameter $\mathbf{k}$ for both figures is set to 0.5, the constant assignment of $\mathbf{w}$ varies according to the values of the abscissa. Due to the relation of the parameters $\mathbf{k}$ and $\mathbf{w}$ the result will shift along the axis of abscissa for a different assignment of $\mathbf{k}$. The best accuracy of the MFN using the output of the classifier based on the facial expressions is obtained by setting the elements of $\mathbf{w}$ to 775, which leads to a ratio of 0.58 correctly classified frames (averaged over the subjects).



(a) Facial expressions ($\mathbf{k}_f = 0.5$).

(b) Prosodic analysis ($\mathbf{k}_p = 0.5$).

(c) Facial expressions, prosodic analysis, gestural analysis ($\mathbf{k}_f = 0.5$, $\mathbf{k}_p = 0.5$, $\mathbf{k}_g = 0.5$)

(d) Facial expressions, prosodic analysis, gestural analysis ($\mathbf{k}_f = 0.5$, $\mathbf{k}_p = 4$, $\mathbf{k}_g = 4$)

Fig. 7. Accuracies (black thick line) and the corresponding standard deviation (gray slope) of different modalities connected to the MFN. The best performing recognition rate is marked by a diamond shaped point. For each plot the parameter of $\mathbf{k}$, which weights the input, is fixed, while the assignment of the lateral smoothness parameter $\mathbf{w}$ is fixed to the value of the abscissa.

The results confirm the low expression level in the facial channel which has already been presumed by the FACS coding performed in the pre-analysis. The prosodic analysis, which suffers from multiple missing decisions due to the silence of the speaker, shows a rather good frame-wise accuracy of 0.75 setting the elements of $\mathbf{w}$ to 775. As to be seen in both figures, there is a broad optimum of good performing parameter assignments.

In case of the gestural analysis, the LOSO cross-validation showed that the occurrences of self-touch has a significant higher prior in the CH event (16%) than the BL event (2%). Since, the absence of self-touch is assumed to give no evidence to any event, these frame have no impact to the final decision. As a result, the gestural analysis is not able to indicate the BL event and, therefore, cannot be utilized for a stand-alone classification of the events.

The recognition performances using all modalities are shown in Fig. 7c and 7d. Fig. 7c shows the performance using uniform weighting of all modalities ($\mathbf{k}_f = 0.5$, $\mathbf{k}_p = 0.5$, $\mathbf{k}_g = 0.5$) reaching an accuracy of up to 0.74, which is slightly worse than the prosodic channel alone. The estimate has also to deal with a higher standard deviation (indicated by the gray slope around the accuracy). The performance can be attributed to the rather low performance of the facial expression channel. Furthermore, a uniform weighting does not necessarily imply that each

modality has the same influence on the re-estimation of the MFN. Each modality has its own fraction of provided decisions (e.g. the classifier based on facial expression provides decisions for all frames, while the prosodic analysis only for 15.9% of the frames and the gestural analysis only for 9% of the frames). Fig. 7d shows the same analysis using a more balanced configuration by the amplifying influence of the prosodic and gestural analysis ($\mathbf{k}_f = 0.5$, $\mathbf{k}_p = 4$, $\mathbf{k}_g = 4$). The accuracy of the new setting achieves up to 0.8 and, therefore, clearly outperforms each single modality.

Table 2. Classification results for 13 individual subjects of the corpus in percent. For prosody, the values of unweighted accuracy with mean and standard deviation are given for ten-fold validation. For facial activities, the accuracies for BL and CL are given.

Subject	Prosody	Facial Activities (BL)	Facial Activities (CH)	Fusion
33	65.22±11.22	91.02	11.33	89.6
34	51.72±4.31	98.44	71.09	95.7
35	75.00±2.14	86.33	66.41	90.2
36	57.14±1.19	71.88	58.20	100.0
46	54.17±3.84	37.50	20.70	89.8
50	60.00±13.41	66.88	43.75	94.5
52	78.95±7.01	81.64	48.44	82.7
54	66.67±2.14	91.02	91.80	62.0
55	68.00±2.40	46.09	5.08	77.5
59	40.00±2.40	57.03	24.61	100.0
61	61.90±1.19	30.86	55.47	71.2
62	41.18±3.47	0.00	50.8	59.8
71	66.67±1.98	95.31	81.25	24.9

The right column in Table 2 shows the frame-wise accuracies in percent compared to the uni-modal classification results using the already presented multi-modal parameter setting. It can be seen that the average accuracy has been significantly improved. Only subject 71 has an accuracy lower than 50 %.

4 Conclusion

Within this work, we presented a multi-modal classifier system to recognize the affective state of persons involved in a Wizard of Oz experiment. The goal of the experimental design aims at discriminating the two classes *Baseline* (BL) and *Challenge* (CH).

The video channel provides high-level features describing facial expressions and self-touch events obtained by the gesture recognition (i.e. hand and face have contact). Based on the facial features a classifier is trained to recognize the events BL and CH. The self-touch events have been used directly as evidence for the class CH (the corresponding hypothesis was consolidated by cross-validation).

A prosodic analysis is performed on the audio channel resulting in a classifier capable of well discriminating between BL and CH. Based on the three input channels a MFN was utilized. The MFN is a powerful approach to fuse decisions of different modalities preserving temporal dependencies. The MFN reconstructs missing decisions which might be present due to the channels characteristics (e.g. speech/non-speech) or different temporal resolution (e.g. physiological features such as respiration describe a larger time scale than prosodic classifiers working on word scale).

Furthermore, we could show that the approach is capable of handling one class evidences given by self-touch events of the gesture recognition. Each channel can be individually weighted according to occurrences and reliability of the decisions provided. Future work will aim at making use of the MFN approach in a real-time classifier system in collaboration with members of the research center (SFB/TRR 62) in order to realize a prototypical companion.

Acknowledgments. This work was supported by the Transregional Collaborative Research Center SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

References

1. Bartlett, M., Littlewort, G., Vural, E., Lee, K., Cetin, M., Ercil, A., Movellan, J.: Data Mining Spontaneous Facial Behavior with Automatic Expression Coding. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 1–20. Springer, Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-70872-8_1
2. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N.: Whodunnit - Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech. Computer Speech and Language 25(1), 4–28 (2011)
3. Bishop, C.M.: Pattern Recognition and Machine Learning. In: Jordan, M., Kleinberg, J., Schölkopf, B. (eds.) Pattern Recognition and Machine Learning, Springer (2006)
4. Cowie, R., Cornelius, R.R.: Describing the Emotional States that are Expressed in Speech. J. on Speech Commun. 40(1-2), 5–32 (2003)
5. Diebel, J., Thrun, S.: An Application of Markov Random Fields to Range Sensing. In: Proc. of Advances in Neural Information Processing Systems (NIPS), vol. 18, pp. 291–298. MIT Press (2006)
6. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologist Press, Palo Alto (1978)
7. Ganchev, T., Fakotakis, N., Kokkinakis, G.: Comparative evaluation of various mfcc implementations on the speaker verification task. In: Proc. of the SPECOM 2005, pp. 191–194 (2005), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.75.8303
8. Glodek, M., Schels, M., Palm, G., Schwenker, F.: Multi-modal fusion based on classification using rejection option and markov fusion network. In: Proceedings of the International Conference on Pattern Recognition (ICPR). IEEE (to appear 2012)

9. Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R., Mogran, N., Bourlard, H., Hermansky, H.: Automatic speech recognition: An auditory perspective. In: Speech Processing in the Auditory System, Springer Handbook of Auditory Research, vol. 18, pp. 309–338. Springer, New York (2004), http://dx.doi.org/10.1007/0-387-21575-1_6, doi:10.1007/0-387-21575-16
10. Kanluan, I., Grimm, M., Kroschel, K.: Audio-visual Emotion Recognition using an Emotion Space Concept. In: Proceedings of the European Signal Processing Conference (EUSIPCO), Lausanne (2008)
11. Kelley, J.F.: An empirical methodology for writing user-friendly natural language computer applications. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 1983, pp. 193–196. ACM, New York (1983), http://doi.acm.org/10.1145/800045.801609
12. Lausberg, H., Kryger, M.: Gestisches Verhalten als Indikator therapeutischer Prozesse in der verbalen Psychotherapie: Zur Funktion der Selbstberührungen und zur Repräsentation von Objektbeziehungen in gestischen Darstellungen. Psychotherapie-Wissenschaft 1(1) (2011), http://www.psychotherapie-wissenschaft.info/index.php/psy-wis/article/view/12
13. Mahmoud, M., Robinson, P.: Interpreting Hand-Over-Face Gestures. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 248–255. Springer, Heidelberg (2011), http://dl.acm.org/citation.cfm?id=2062850.2062879
14. Metallinou, A., Lee, S., Narayanan, S.: Audio-visual Emotion Recognition using Gaussian Mixture Models for Face and Voice. In: Proc. of the IEEE Int. Symposium on Multimedia, Berkeley, CA, pp. 250–257 (December 2008)
15. Niese, R., Al-Hamadi, A., Panning, A., Brammen, D., Ebmeyer, U., Michaelis, B.: Towards pain recognition in Post-Operative phases using 3D-based features from video and support vector machines. International Journal of Digital Content Technology and its Applications (2009), http://www.aicit.org/JDCTA/paper_detail.html?q=92
16. Paleari, M., Huet, B., Chellali, R.: Towards Multimodal Emotion Recognition: A new Approach. In: Proceedings of the ACM International Conference on Image and Video Retrieval, Xi'an, China, July 5-7 (2010)
17. Palm, G., Glodek, M.: Towards emotion recognition in human computer interaction. In: Proceedings of the Italian Workshop on Neural Networks WIRN (to appear, 2012)
18. Panning, A., Al-Hamadi, A., Michaelis, B.: Active Shape Models on Adaptively Refined Mouth Emphasizing Color Images. In: WSCG Communication Papers, pp. 221–228 (2010)
19. Panning, A., Siegert, I., Al-Hamadi, A., Wendemuth, A., Rösner, D., Frommer, J., Krell, G., Michaelis, B.: Multimodal Affect Recognition in Spontaneous HCI Environment. In: IEEE International Conference on Signal Processing, Communications and Computings, ICPCC 2012 (to appear, 2012)
20. Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., Otto, M.: LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions. In: Calzolari (Conference Chair), N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proc. of the Eighth International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA) (May 2012)

21. Saeed, A., Niese, R., Al-Hamadi, A., Panning, A.: Hand-face-touch Measure: a Cue for Human Behavior Analysis. In: IEEE Int. Conf. on Intelligent Computing and Intelligent Systems, vol. 3, pp. 605–609 (2011)
22. Saeed, A., Niese, R., Al-Hamadi, A., Panning, A.: Hand-face-touch measure: a cue for human behavior analysis. In: 2011 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), vol. 3, pp. 605–609 (2011)
23. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.: Acoustic emotion recognition: A benchmark comparison of performances. In: Proc. of IEEE Workshop on Automatic Speech Recognition Understanding (ASRU), Merano, Italy, pp. 552–557 (December 2009)
24. Schuller, B., Valstar, M.F., Eyben, F., McKeown, G., Cowie, R., Pantic, M.: AVEC 2011–The First International Audio/Visual Emotion Challenge. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part II. LNCS, vol. 6975, pp. 415–424. Springer, Heidelberg (2011)
25. Siegert, I., Böck, R., Philippou-Hübner, D., Vlasenko, B., Wendemuth, A.: Appropriate Emotional Labeling of Non-acted Speech Using Basic Emotions, Geneva Emotion Wheel and Self Assessment Manikins. In: Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2011, Barcelona, Spain (2011)
26. Soleymani, M., Pantic, M., Pun, T.: Multi-Modal Emotion Recognition in Response to Videos. IEEE Transactions on Affective Computing 99, Preprints (November 2011) (in press)
27. Vural, E., Çetin, M., Erçil, A., Littlewort, G., Bartlett, M., Movellan, J.: Machine learning systems for detecting driver drowsiness. In: Takeda, K., Erdogan, H., Hansen, J.H.L., Abut, H. (eds.) In-Vehicle Corpus and Signal Processing for Driver Behavior, pp. 97–110. Springer, US (2009), http://dx.doi.org/10.1007/978-0-387-79582-9_8
28. Wagner, J., Lingenfelser, F., André, E., Kim, J.: Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data. IEEE Trans. on Affective Computing 99, Preprints (2011)
29. Wendemuth, A., Biundo, S.: A Companion Technology for Cognitive Technical Systems. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) COST 2102. LNCS, vol. 7403, pp. 89–103. Springer, Heidelberg (2012)
30. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control.. Clinical Neurophysiology 113(6), 767–791 (2002), http://view.ncbi.nlm.nih.gov/pubmed/12048038
31. Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., Freeman, W.T.: Eulerian video magnification for revealing subtle changes in the world. ACM Trans. Graph (Proceedings SIGGRAPH 2012) 31(4) (2012)
32. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book (for HTK Version 3.4). Cambridge University Engineering Department, Cambridge, UK (2006), http://nesl.ee.ucla.edu/projects/ibadge/docs/ASR/htk/htkbook.pdf
33. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(1), 39–58 (2009)
34. Zeng, Z., Tu, J., Pianfetti, B., Huang, T.: Audio-visual Affective Expression Recognition through Multi-stream Fused HMM. IEEE Trans. on Multimedia 4, 570–577 (2008)

Author Index