INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin

www.informatik2011.de

Social Signal Processing in Companion Systems - Challenges Ahead

Georg Layher[1], Stephan Tschechne[1], Stefan Scherer[1]
Tobias Brosch[1], Cristóbal Curio[2], Heiko Neumann[1]

[1] Institute of Neural Information Processing, University of Ulm, Germany,
{georg.layher, stefan.scherer, stephan.tschechne, tobias.brosch, heiko.neumann}@uni-ulm.de
[2] Max-Planck-Inst. for Biological Cybernetics, Tübingen, Germany,
cristobal.curio@tuebingen.mpg.de

Abstract: Companion technologies aim at developing sustained long-term relationships by employing emotional, nonverbal communication skills and empathy. One of the main challenges is to equip such companions with human-like abilities to reliably detect and analyze social signals. In this proposal, we focus our investigation on the modeling of visual processing mechanisms, since evidence in literature suggests that nonverbal interaction plays a key role in steering, controlling and maintaining social interaction between humans. We seek to transfer fragments of this competence to the domain of human computer interaction. Some core computational mechanisms of extracting and analyzing nonverbal signals are presented, enabling virtual agents to create socially competent response behaviors.

1 Introduction and Motivation

The development of future companion technologies necessitates creating agents that are personalized in terms of building relationships with their (human) interlocutors on the basis of intelligent multimodal interaction. A key functionality of artifacts which function as true companions is their ability to build human-computer relationships on a long-term scale [BP05]. The enabling technology to develop such skills for artificial companions is captured in the five-star model by [BM08] which incorporates contributions along several dimensions such as utility, the form of the agent and its interactive capabilities, social attitudes of the companion's role and persuasion, emotional skills such as expression and empathy, and aspects of personality and trust. Such artificial companions need to be sensitive to emotion and disposition in vision and speech, gesture, touch, and individuality concerning autonomy and personality. In communication processes nonverbal social signaling conveys determination, interest, relatedness, etc. Due to variations and subtleties in expression these are not easy to analyze automatically from real data and in different contexts. In order to detect social signals the features and signal properties must be defined and then detection and interpretation mechanisms need to be developed [Pen07].

In this paper, we focus our investigation on the modeling of visual processing mechanisms to acquire and analyze socially communicative signals since evidence suggests

INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin

www.informatik2011.de

that nonverbal interaction plays a key role in steering and controlling social interactions [BP05, FW04]. This focus, however, does not deny the importance and central role that verbal communication plays in isolation and in conjunction with nonverbal signals in natural communication, specifically in task-oriented interaction. Below, we will briefly summarize previous work on the processing and analysis of social signals to augment nonverbal social interaction frameworks. Based on this work we sketch our approach and outline a roadmap to develop key mechanisms for visual analysis of social signals and behavior.

2   Social Signal Analysis in Relational Nonverbal Behavior

Future companion technologies need to be equipped with facilities that allow constructing relationships and to elaborate, maintain and evaluate them on a long-term scale. The range of functions and applications for companion relationship is multi-faceted and operates upon different sensory modalities and levels of cognition and relationship modeling [BP05]. In order to achieve companion functionality the capabilities of an agent must be so that it has extended verbal and nonverbal communication skills to enable the management of personal relationship and for the perpetuation of communication. Unlike classical dialog understanding based on speech recognition methods the topic of social interaction signaling and analysis provides an alternative and extended framework of discourse [Pen07]. Several authors have begun to investigate the different codes, the related behavioral cues and the functions of social signals [VPBP08]. Nonverbal communication aggregates behavioral cues (e.g., posture, gesture, facial expression) to serve different functions, such as to signal attention, managing social contact and interaction, form messages of agreement (or disagreement) for continuation or disconnection, expressing empathy, etc. [AR92]. The display of interpersonal attitudes through the use of immediacy behaviors (expressed by direct body and facial orientation in close distance), direct gaze, frequent gesturing, open posture, smile and pleasant facial expression and animation, etc., signal active interest and involvement in an interaction at various levels [Arg88]. An automatic analysis of communication skills from realistic data must provide a battery of mechanisms and their coordinated fusion (at various levels) which allows to instantiate a rich repertoire of detecting and interpreting social signals. Social signal processing has been investigated for different classes of nonverbal behaviors and the functions they support. The automatic analysis of nonverbal cues can be distinguished into approaches of signal processing and analysis from a third-person as well as approaches that operate from a first-person perspective. Examples of the former are pioneering works that have analyzed temporal periods in negotiation scenarios and their outcome prediction, the analysis of small group interactions, including the detection of roles of different participants and the collective actions in social groups (see [VPBP08] for an overview). An analysis of cues from a first-person perspective has to cope with the problem that the artificial agent itself takes a role of a (possibly human-like) social actor. Numerous contributions have been published that were mainly developed in restricted settings and scenarios, such as, e.g., for automatic head pose and gaze detection using different sensors, facial expression analysis, and human body tracking [WADP97, EP97, MCT09]. The development of mechanisms that seem-

INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin

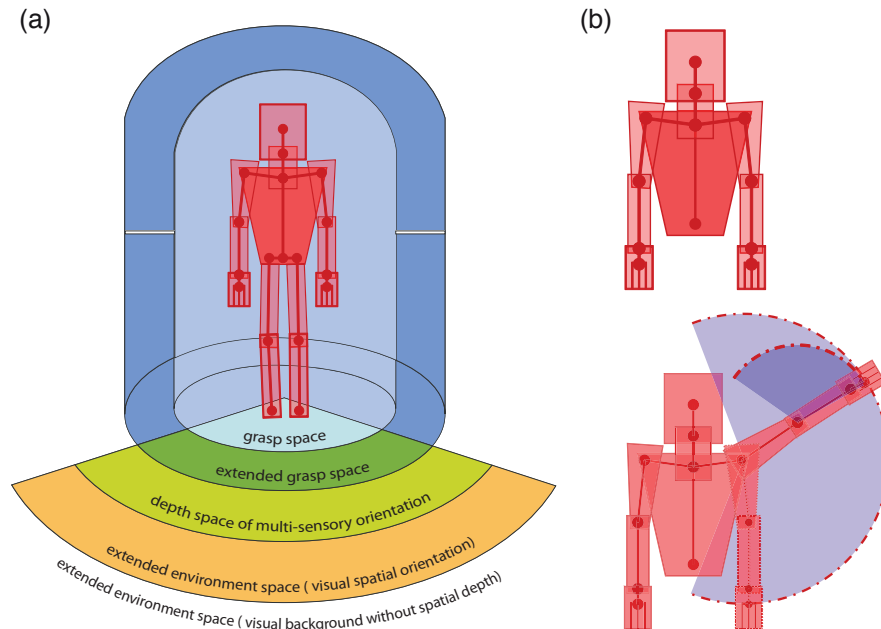www.informatik2011.de

(a)   (b)

Figure 1: Bodily dimensions and their relevance for the processing of nonverbal social signals. In (a), the extra-personal space of an agent is shown (modified after [Grü78]). From the center to the periphery, the extra-personal space can be partitioned into four regions: the immediate space of reaching, an extended grasp space, a depth space of multi-sensory orientation and an extended environment space. Each of the regions entails different requirements and constraints for the sensory processing. (b) shows a depiction of an eight-joint human upper torso model for articulated motion. The configuration is used as an outline of the proposed computational mechanisms for detecting visual input and analyzing encoded social signals. Head and torso are joined through the neck which separates the two components. The upper arm segments are joined to the torso allowing the generation of various gestures (two exemplary configurations shown). Each arm is configured as a chain of upper arm and lower arm segment, and the hand joint to the lower end of the lower arm. It should be noted that we do not follow a full 3D modeling approach to solve the kinematics of this body model. This configuration, however, helps to guide the subsequent visual processing mechanisms. Depiction of the range within reach of an arm with upper arm, lower arm, and hand (bottom). Based on the angular degrees of freedom at the joints of the individual limbs the position of each component can be constrained in a sequentialized detection process (see text for discussion).

ingly fit into a cognitive architecture of social signal analysis is still in its infancy. The challenges include the use of real-world data gathered from everyday actions and the analysis of multi-cue signals over different time spans in order to reduce inherent ambiguities. Recent studies using embodied relational agents lack mechanisms for automatic user affect analysis [VPBP08, BP05]. Our own work, as outlined in the next section, belongs to the second group of work that analyzes social input signals from a first-person perspective. The aim in the long run is to equip the computer with human-like abilities to seamlessly

fit into a scenario where the computer acts as a partner in complex human-computer inter-
actions as if they were human-human.

3   Modeling Framework for First-Person Social Signal Analysis

The aim of this section is to outline the generic principles of our approach to the automatic
analysis of visual input data that is acquired in a social perception scenario. We pursue
a biologically inspired approach of developing visual, perceptual and cognitive abilities
adopting known principles of the functional organization in the primate brain. The goal
is to develop an architecture as integral part of an artificial companion which is capa-
ble of analyzing social signals from a first-person perspective. It is intended to describe
few design principles for the analysis of complex spatio-temporal action or activity se-
quences. These results will then be incorporated into computational vision approaches for
the processing of parallel stereoscopic video input data. The approach is motivated by
the suggestion that the detailed study of elaborated mechanisms, flows and protocols in
human-human interaction helps to identify the rules for developing skillful future human-
machine communication [PK02]. In support of the suggested approach there exists a rich
body of experimental findings indicating that primary feature processing in the primate
cortex is organized along segregated pathways of form and motion processing, namely the
ventral and the dorsal streams, respectively. They converge to build representations in the
superior temporal sulcus (STS) with cells showing a variety of selectivities to head, body,
and hand poses and actions as well as of their conjunction. This suggests that these cells
contribute to build up functional networks to instantiate the components of a social brain
[Bro96, PHB$^+$89].

The analysis of human pose and gestures from video streams focuses on articulated move-
ments of single or multiple actors in the scene. Movement patterns in this case arise
from components (torso, limbs, etc.) in possibly different directions and speeds, but with
joint components that constrain the flow interpretation with common joint movements
[DSK11]. While many approaches follow a holistic approach to analyzing object move-
ments, we suggest an actively guided approach to sequential segmentation and analysis
of body parts. The processing stages depend on the prior results and expectations de-
rived thereof, which is advantageous to reduce the search space in analyzing an actor's
articulations. As a reference we utilize an eight-joint body model that captures essential
components of an upper body together with the degrees of freedom of the limbs and their
configurations to signal subtle social signals during conversation (see Fig. 1 (b)). In a nut-
shell, we argue that the analysis could be driven by selectively detecting body parts, like
the head or face, which then serve as a starting point for the subsequent detection of, e.g.,
upper torso and upper arm limbs. Unlike, e.g., [KDHU10], we anchor certain positional
priors (with parameterized orientation) for individual body parts to initiate the analysis of
other object components. The individual priors are designed to trigger visual routines op-
erating upon incremental representations of task related processing [Ull84]. The suggested
operational principles will be outlined and further detailed below.

3.1 Visual Processing and the Encoding of Social Signals

In order to build a base representation for analyzing various possible scenic events and the possible configurations of objects and object parts a rich set of necessary features needs to be detected from input data streams. Such features may be used for different computations in order to generate the required internal sensory representations. Different regions of extra-personal space of an agent can be identified which, in turn, impose different constraints upon the sensory processing. In an earlier work [Grü78] distinguished four such regions, namely (i) the immediate space of reaching, grasping and manipulation, (ii) an extended grasp space (that can be explored by manipulatory tool augmentation), (iii) a depth space of multi-sensory orientation, and (iv) an extended environment space composed of figural objects and visual background (see Fig. 1 (a)). It is evident that some generic visual features maybe useful to visually analyze all these spatial regions, whereas more focused and task-related processing may be seen for processes that selectively operate upon different spatial regions. For example, we can covertly monitor activities in the *far space* to judge potential threat or communicative attempts. Here, only (relative) motion information is necessary to selectively compute relevant information. When another actor approaches distinct brain areas show increased activation to monitor his/her potential intentions [MPM05]. In the transition range between *far and near space* head and body appearance in visual scenes needs to be reliably detected. An unsolved problem in general vision-guided mechanisms is the proper handling of mutual occlusions, either of distinct objects or of multiple body parts, as they occur for unconstrained body and arm postures. We utilize a cortex-like hierarchical processing model for initial static feature detection and subsequent aggregation for mid-level feature detection for static form features. In parallel, we extract motion features and integrate them to disambiguate and resolve initial uncertainties in the spatio-temporal configuration patterns. Intermediate level complex motion patterns and temporally filtered response configurations help to extract motion information at an intermediate-level scale. In the *near space*, in which grasping and instrumented manipulation takes place, nonverbal visual interaction is dominated by face-to-face communication and selective cueing to target objects of potential joint interest. New codes and behavioral cues thus need to be analyzed. For example, face-to-face communication relies on the detection of postural congruence (almost frontal body and head-to-head poses) as well as mutual gaze. The initial stages of model cortex filtering provide input for further grouping and competition to feed into a scheme of unsupervised learning of, e.g., head poses, which then could be used for estimating pose categories [WN08]. Also representations of eye gaze direction could be generated by using filtering and relative phase responsiveness derived from the pupil and sclera pattern in human eyes [LWB00]. Taken together, the initial stages of form and motion processing together with stereoscopic correspondences derived from image pair matching provide a rich set of features to serve the analysis of social signals.

INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin

www.informatik2011.de

3.2 Low-Level Visual Mechanisms for Basic Form and Motion Feature Detection and Grouping

Initial processing for static form feature extraction is based on a modified variant of the biologically inspired object-recognition model proposed by [ML08] (see also [RP99, SWP05]). In a nutshell, the model architecture consists of a processing hierarchy of stages consisting of alternating levels of filtering and selection (pooling) steps, which start at the level of the primary visual cortex, or V1. These operate at different scales of spatial neighborhood. The non-linear pooling over a lateral neighborhood aims at achieving input pattern invariance against variations in size, rotation and position. Our model consists of five different processing stages. An input image is transformed into a pyramidal representation of different spatial scales. Each scale is convolved with 2D Gabor filters in four orientations, resulting in four orientation fields for each of the three different scales (spatial frequency selectivity). In each scale, the pooling of neighboring filter responses achieves tolerances against variations in location, shape position and size. In the next successive processing stage, intermediate level features are learned by selecting the most descriptive and discriminative prototypes among an exhaustive number of response patches by random sampling in the response distribution of the previous processing stage. The resulting prototype patterns denote filters with complex feature selectivities topographically organized around the spatial locations of their most likely occurrence. Again, their responses are pooled afterwards to gain an increased property of invariance. At the final stage, responses from all prototypical complex filters are integrated over positions and scales by using a winner-take-all strategy selectively operating on the different spatial scales. The responses are combined into a single feature vector, which serves as input to a linear SVM. The SVM finally allows to classify the individual feature object representations (for details, see [ML08]).

Processing along the motion pathway is achieved by mainly two stages. At the early stage of primary visual cortex, input stimuli are analyzed in parallel for movements along different directions. These initial direction responses are calculated using an extended scheme of frame-to-frame correlation matching. This matching process is accelerated significantly by utilizing a variant of the Census transform [Ste04]. Model MT is the next stage in the processing hierarchy where cells with increased receptive field sizes integrate initial responses from correlation detectors in model V1. Model MT cells, in turn, send top-down feedback signals to modulate initial responses and thus stabilize the motion detection and integration process (for details, see [BN07, BTKN11].

The resulting representations in the form and the motion pathway serve as base codes for subsequent processing stages that are task-driven in the context of analyzing social signals and their prerequisites. Several mechanisms are outlined in the following.

INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin

www.informatik2011.de

3.3 Mid-Level Processing Mechanisms

We have motivated to employ a task-driven processing mechanism using visual routines in order to selectively constrain the search space to analyze possible shape configurations. For that reason, we suggest several intermediate-level mechanisms which operate upon the base level representations to generate incremental encodings to support task-related purposes. We reliably estimate the head pose in two processing steps. In the first step, four different facial features are detected and localized in the images of a stereo pair. This is accomplished by employing the hierarchical shape processing architecture outlined above and the selection of intermediate-level facial features along the hierarchy. These enable the detection of the eyes and the mouth corners within the images of a stereo pair. Intermediate-level features are more selective but still have a topographic localization. Such features have been analyzed and shown to be superior in terms of their specificity and their relative frequency in comparison to low-level features [UVNS02]. For stereoscopic matching and pose estimation this property has the advantage of reducing the false target candidates in stereo matching but at the same time allow to estimate disparities such that a proper depth resolution is preserved ([LLN$^+$11]). In the second step, after successful localization of intermediate level facial features, the associated disparities are determined by maximizing the correlation of a feature in the left image and its counterpart in the right image within a local neighborhood. Due to the task-related processing strategy we do not need to estimate a dense disparity map which reduces computational costs considerably. Given the disparity values as well as the focal length and the baseline of the stereo camera system, the 3D world coordinates and depth values of the facial features can be calculated. The orientation of the head is then estimated by fitting a plane (facial plane) through the four facial feature positions located in space (see Fig. 2 (a) and (b)). Once the head and
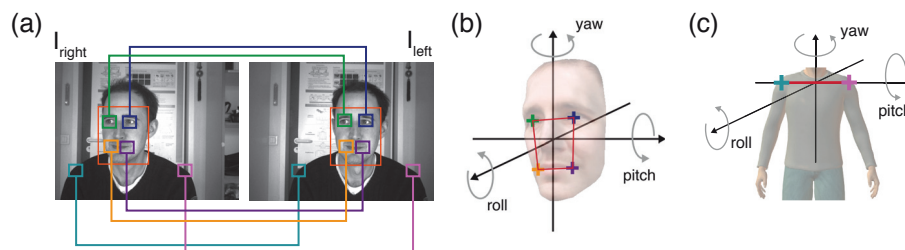

Figure 2: Estimation of the head and body pose. Building upon the position of the face, four facial features, as well as the two shoulders are localized in a stereo image pair using features of intermediate level complexity (a). The disparities of the features are determined by maximizing the correlation between a feature in the left image and its counterpart in the right image within a local neighborhood. In case of the head, pose is estimated by fitting a facial plane onto the inferred 3D surface information of the facial features. The orientation of the resulting plane is then used as an direct estimate of the head orientation (b). The procedure for the estimation of the upper torso orientation is almost identical, yet using the orientation of the connecting line between the two shoulders as a representative of the orientation of the torso (c).

its orientation have been detected the search for the expected regions of the upper torso as well as the upper arm joints and shoulder regions can be confined to regions defined by anthropometrically driven expectations. Stereo matching of regions with shoulder appearances utilizing intermediate level features allow computing a rough estimate of the spatial orientation of the body torso like the head pose estimation outlined above (see Fig. 2 (c)).

One of the most challenging problems is the detection and segmentation of articulated arms and their movement [WH99, Tur04]. Here we investigate the segmentation of the upper arm and its image orientation. Initial contrast detection is accomplished by using the processing cascade as outlined above. Oriented contrasts can be subsequently grouped to form extended boundary signals [WN09] followed by the estimation of a symmetry axes. These static features are combined with evidences derived from motion and spatio-temporal occlusion signals. The detection of spatio-temporal occlusion from motion is motivated by Gibson's and colleagues' observations [GKRW69] showing that the pattern of deletion and accretion of optical texture provides evidence for the presence of mutual surface occlusions. We employ here a computational mechanism that is based on the detection of discontinuities in the initially detected and integrated motion patterns. Motion discontinuities are combined with responses from the detection of temporal changes in motion energy which encode occlusion and dis-occlusion regions in the motion field [BON08]. An extension of these basic processing mechanisms has been proposed in [TN11] to allow the segregation of figural surfaces from structured background. The architecture is reminiscent of the one proposed by [CSNvdH07] to model figure-ground segregation in *stationary* scenes. We employ this motion-based mechanism for segregating the arm limbs and hands from background and also to segregate bodies at full scale when they move in the scene or approach the monitoring observer.

Many biological forms share the property of highly symmetrical structure and appearance. This also holds for the upper and lower arm segments. The appearance of such surface patches can be compactly described by a symmetry axis which, in turn, can be determined by a medial axis transform [Blu67]. Here we incorporate the approach developed by Curio and coworkers [EST+09] to compute medial features from grey level input images. The algorithm consists of a two-stage process. First, a vector field of diffusion flow, emanating from local contrast boundaries, is computed through energy minimization that regularizes the simultaneous approximation and smoothing of a gradual activation surface given the image gradients [XP98]. Second, local sinks in the resulting vector fields are detected that depict local nodes on the medial axis. It is worth mentioning that unlike stick model approaches (e.g., [OTA08]), we make use of a richer repertoire of input features, namely the outer rim as well as the symmetry axis. This has the advantage to fuse the form information with motion and occlusion/dis-occlusion information. The goal of fusing several available visual input channels is to increase reliability by building up a rich set of powerful visual features through the convergence of ventral (shape) and dorsal (motion) stream representations. The estimation of the orientation of the upper arm limb (at the shoulder joint) is indicative for the potential locations of the lower arm and the hand, based on the degrees of freedom at elbows and wrists (see Fig. 1 (b)). Consider the configurational space of the upper and lower arm and the hand which is approximated by the appearance of a half-circular region (large gray circular arc region). Its estimated orientation for the upper arm

INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin

www.informatik2011.de

using the approach sketched above reduces the possible occurrences of the lower arm and hand. Based on anthropometric constraints as well as the degrees of freedom of the limbs the lower arm and hand is bound to the circular sector defined by the body and the upper arm axes (small blue circular arc region). We emphasize here that this discussion focuses on the great circle of the spatial hemisphere of arm reaches in the coronal plane. Even for manual operations in front of an actor, the detection of image appearances of the lower arm and hand is still constrained to the regions outlined in Fig. 1 (b) (bottom). These outer limb components can be detected by a further stage of symmetry-based detection (using combined medial features and motion information) as well as an active segmentation component, such as proposed by Aloimonos and colleagues [AGFO10]. The latter component is briefly outlined below.

3.4 Active Segmentation of Specific Body Parts

The robust segmentation of the body limbs, e.g. upper and lower arms, hand and even individual fingers convey further information about, e.g., exposure, gestures, self-presentation, and conversational distance [Arg88]. For real scenes, as outlined above, we suggest that the visual extraction of arm and hand poses is organized sequentially by an attention-guided search process which proceeds in a coarse-to-fine manner. Segmentation of the hand is triggered by higher-order visual routines which operate to build an incremental representation providing a link to sensory-motor tasks [Ull84]. In the context of lower limbs and hands detection we seek evidence for the presence of the upper arm segment as outlined above. A target region in the occurrence sector of lower arm limb and hand can be identified after the upper arm has been detected (compare Fig. 1 (b)). We utilize the approach of [MA09]. In a nutshell, the algorithm actively centers the local reference coordinate system at a selected target region (simulating an artificial saccadic eye movement). Using a space-variant image representation that centers the high-resolution at the gaze center leads to a simplified segregation of the focused target region from the current background (that surrounds the current target). The space-variant imaging is reminiscent of the foveation of the human eye and the non-linear transform of the input into a cortical representation. A binary graph cut algorithm (e.g., [BK04]) segregates the figural segment against the peripheral background using the figure boundary to steer the min-cut segmentation. The segmented target region can be tracked over time to estimate characteristic temporal signatures in nonverbal communication.

4   Results

In the following, we briefly show some results of the components in the proposed architecture, that have been implemented and tested in real world scenarios. First, the capability of the head and body pose estimation approach described in Section 3.3 is shown. Second, the essential processing stages of the mechanisms responsible for the extraction of a skeletal representation of the upper body are shown using the example of the forearm.

INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin
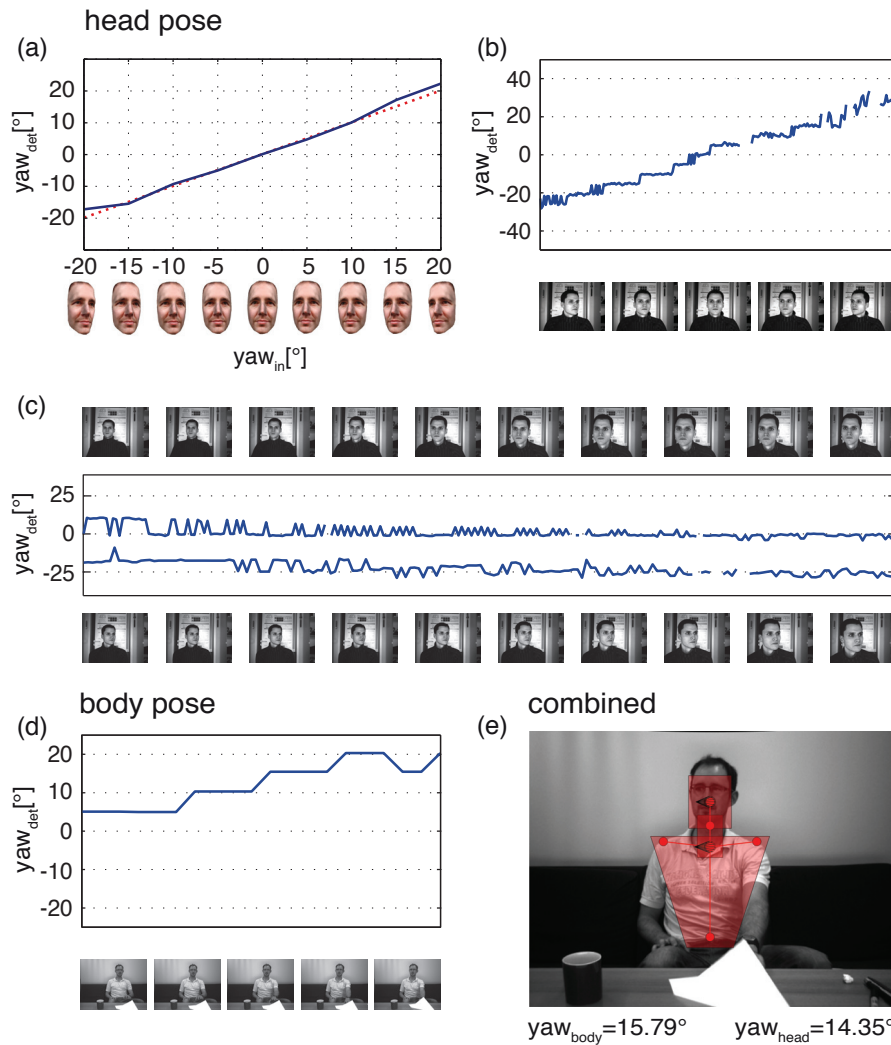
www.informatik2011.de

Figure 3: Stereoscopic head pose ((a)-(c) and (e)) and body pose estimation ((d) and (e)). (a) Artificial stereo images with known ground truth were used to evaluate the precision of the head pose estimation. The error for horizontal head poses remained below $3°$ (with larger yaw angles producing larger errors). (b) A stereo camera system was used to test the proposed head pose estimation approach under real world conditions. The actor was instructed to rotate his head systematically from the left to the right. Even though no ground truth data is available, it can be seen that the estimated head pose reflects the course of the actor's head orientation. (c) The capability of the proposed approach under varying camera-actor distances was tested using two real world sequences with different but constant yaw angles (above and below the plot). As it can be seen, there is only a little variation in the estimation quality for the head poses at different distances. (d) Estimated orientation of the upper torso using a sequence of real world images. (e) Estimated orientation of the head and the upper torso in combination with the inferred configuration of the affected joints.

INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin

www.informatik2011.de

4.1 Head and body pose estimation

We evaluated the proposed stereoscopic head pose estimation approach under two different conditions. First, we used an in-house head pose database to generate test data with known ground truth. In Fig. 3 (a), the estimated head pose is shown. As one can see, the estimation error increases for larger yaw angles, but never exceeds $3°$ over head poses in the range of $[-20°,20°]$. Second, a sequence of real world images was used. The actual head pose within the sequence was unknown, but the subject was told to systematically rotate his head from the left to the right. As shown in Fig. 3 (b), the estimated head pose reflects that fact. One of the key features of the proposed approach, namely its invariance against varying camera-actor distances, is shown in Fig. 3 (c). As described in Section 3.3, we used an almost identical approach to estimate the orientation of the upper torso. Fig. 3 (d) shows the estimated yaw angle for a real world image sequence. A combined estimation result for the head and the upper torso orientation is shown Fig. 3 (e). The joints for the head and the shoulders were localized automatically, whereas the remaining three joints were inferred via anthropometric proportions. It is worth noting, that the underlying classifiers used for the localization of the head and the facial features were all trained using the FERET Database [PMRR00] and thus are independent of the test data used here.

4.2 Extraction of skeletal representations of upper body parts

Fig. 4 shows results for combination of medial axis transform and motion-dependent figure-ground segregation (as described in Section 3.3). In the illustrated sequence, a person is waving its arm up and down. The medial axis transform is computed by using the luminance channel only. This produces, apart from the desired structural cues inside the moving limb many undesired structural cues, like those between arm and head, which might impair further interpretation tasks. The medial axis representation alone does not allow a distinction between relevant and irrelevant axes.

However, the scenic motion incorporates cues about scene segmentation, figure-ground organization and border ownership. In Fig. 4 the bottom row illustrates how a segmentation into foreground and background is achieved. First, optic flow is computed, showing speed and direction of moving parts in the image. Spatio-temporal filtering yields signals for occlusions and dis-occlusions. The border ownership signals at regions of (dis-) occlusion indicate which surface currently "owns" the attached image region and thus the associated surface boundary. Together, this information allows to separate the figure foreground from the background. In the illustration, ownership is indicated with a directional color code. The juxtaposed organization of occlusion and dis-occlusion regions allows a highlighting of moving surfaces that are in the foreground.

A linear combination of the resulting signal responses of the medial axis transform with the foreground signal achieved from motion processing yields a significant reduction of axis features. Now, axes can be classified to belong to objects in the foreground, whereas others can be ignored in further processing steps.

INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin
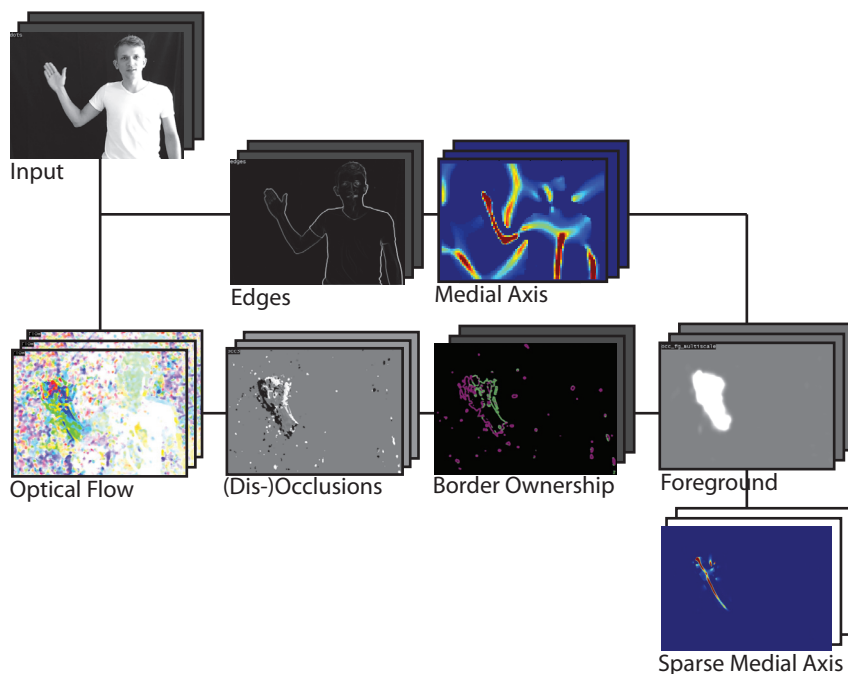
www.informatik2011.de

Figure 4: Medial axis transform and figure-ground segregation. Top: The medial axis transform finds symmetry features that coincide with significant object structures, but also produces many undesired responses. Bottom: Processing of motion signals allows segregation of moving objects in foreground. Linear combination of both results yields a sparse representation of axes that have a high importance for structural scene interpretation.

5 Brief Summary and Outlook

We have presented core computational mechanisms of extracting and analyzing nonverbal social signals to enable an agent creating socially competent response behaviors. The approach is motivated by the suggestion that the detailed study of elaborated mechanisms, flows and protocols in human-human interaction helps to identify the rules for developing skillful future human-machine communication [PK02]. Along the generic architecture we have demonstrated the capabilities to process head and body postures in parallel to skeletal representations of upper body parts, using advanced biologically motivated mechanisms of form and motion processing. Considered in isolation, these mechanisms are already capable of giving cues on particular aspects of nonverbal signals (such as the attentiveness) but need to be further integrated and combined to allow a richer and more meaningful interpretation of social signals. For instance, they provide input for dynamic scene segmentation e.g. to localize the hands and compute their temporal signature (not shown here).

INFORMATIK 2011 - Informatik schafft Communities
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin

www.informatik2011.de

References

[AGFO10]   Y. Aloimonos, G. Guerra-Filho, and A. Ogale. The Language of Action: A New Tool for Human-Centric Interfaces. *Human Centric Interfaces for Ambient Intelligence. H. Aghajan, J. Augusto, and R. Delgado (Eds.), Elsevier*, pages 95–131, 2010.

[AR92]   N. Ambady and R. Rosenthal. Thin Slices of Expressive behavior as Predictors of Interpersonal Consequence : A Meta-Analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[Arg88]   M. Argyle. *Bodily Communication*, volume 2nd. Methuen, 1988.

[BK04]   Y. Boykov and V. Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[Blu67]   H. Blum. A Transformation for Extracting New Descriptors of Shape. *Models for the Perception of Speech and Visual Form*, pages 362–380, 1967.

[BM08]   D. Benyon and O. Mival. Landscaping Personification Technologies: From Interactions to Relationships. In *CHI '08 extended abstracts on Human factors in computing systems*, CHI EA '08, pages 3657–3662, New York, NY, USA, 2008. ACM.

[BN07]   P. Bayerl and H. Neumann. A Fast Biologically Inspired Algorithm for Recurrent Motion Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):246 –260, 2007.

[BON08]   C. Beck, T. Ognibeni, and H. Neumann. Object Segmentation from Motion Discontinuities and Temporal Occlusions - A Biologically Inspired Model. *PLoS ONE*, 3:e3807, 2008.

[BP05]   T. W. Bickmore and R. W. Picard. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Transactions Computer-Human Interaction*, 12:293–327, 2005.

[Bro96]   L. Brothers. Brain Mechanisms of Social Cognition. *Journal of Psychopharmacology*, 10(1), 1996.

[BTKN11]   J. D. Bouecke, É. Tlapale, P. Kornprobst, and H. Neumann. Neural Mechanisms of Motion Detection, Integration, and Segregation: From Biology to Artificial Image Processing Systems. *EURASIP J. Adv. Sig. Proc.*, 2011.

[CSNvdH07]   E. Craft, H. Schütze, E. Niebur, and R. von der Heydt. A Neural Model of Figure-Ground Organization. *Journal of Neurophysiology*, 97(6):4310–4326, 2007.

INFORMATIK 2011 - Informatik schafft Communities
www.informatik2011.de
41. Jahrestagung der Gesellschaft für Informatik , 4.-7.10.2011, Berlin

[DSK11]   A. Datta, Y. Sheikh, and T. Kanade. Linearized Motion Estimation for Articulated
            Planes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):780–
            793, 2011.

[EP97]   I. A. Essa and A. P. Pentland. Coding, Analysis, Interpretation, and Recognition of
            Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,
            19:757–763, 1997.

[EST+09]   D. Engel, L. Spinello, R. Triebel, R. Siegwart, H. Bülthoff, and C. Curio. Medial Fea-
            tures for Superpixel Segmentation. In *Proceedings of the Eleventh IAPR Conference
            on Machine Vision Applications (MVA 2009)*, pages 248–252, 5 2009.

[FW04]   C. D. Frith and D. M. Wolpert. *The neuroscience of social interaction: Decoding,
            imitating, and influencing the actions of others*. Oxford University Press, 2004.

[GKRW69]   J. Gibson, G. Kaplan, H. Reynolds, and K. Wheeler. The Change from Visible to
            Invisible: A Study of Optical Transitions. *Perception &; Psychophysics*, 5:113–116,
            1969.

[Grü78]   O.-J. Grüsser. Grundlagen der neuronalen Informationsverarbeitung in den Sinnesor-
            ganen und im Gehirn. In *GI - 8. Jahrestagung*, pages 234–273, London, UK, 1978.
            Springer-Verlag.

[KDHU10]   L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The Chains Model for Detect-
            ing Parts by their Context. In *Proceedings to the twenty-third IEEE Conference on
            Computer Vision and Pattern Recognition, CVPR 2010*, pages 25–32, 2010.

[LLN+11]   G. Layher, H. Liebau, R. Niese, A. Al-Hamadi, B. Michaelis, and H. Neumann. Ro-
            bust Stereoscopic Head Pose Estimation in Human-Computer Interaction and a Uni-
            fied Evaluation Framework. In *to appear in 16th International Conference on Image
            Analysis and Processing (ICIAP'11)*. Springer, 2011.

[LWB00]   S. R. Langton, R. J. Watt, and I. Bruce. Do the Eyes Have it? Cues to the Direction
            of Social Attention. *Trends Cogn Sci*, 4(2):50–59, 2000.

[MA09]   A. K. Mishra and Y. Aloimonos. Active Segmentation. *I. J. Humanoid Robotics*,
            6(3):361–386, 2009.

[MCT09]   E. Murphy-Chutorian and M. M. Trivedi. Head Pose Estimation in Computer Vision:
            A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:607–
            626, 2009.

[ML08]   J. Mutch and D. G. Lowe. Object Class Recognition and Localization Using Sparse
            Features with Limited Receptive Fields. *Int. J. Comput. Vision*, 80(1):45–57, 2008.

[MPM05]   J. P. Morris, K. A. Pelphrey, and G. Mccarthy. Regional Brain Activation Evoked
            When Approaching a Virtual Human on a Virtual Walk. *J. Cognitive Neuroscience*,
            17:1744–1752, 2005.

[OTA08]   K. Onishi, T. Takiguchi, and Y. Ariki. 3D Human Posture Estimation Using the HOG
            Features from Monocular Image. In *Proc. 19th Int.l Conf. on Pattern Recognition
            (ICPR08)*, 2008.

[Pen07]   A. Pentland. Social Signal Processing. *Signal Processing Magazine, IEEE*,
            24(4):108–111, 2007.

[PHB+89]   D. I. Perrett, M. H. Harries, R. Bevan, S. Thomas, P. J. Benson, A. J. Mistlin, A. J. Chttty, J. K. Hietanen, and Fife Ky Ju. Frameworks of Analysis for the Neural Representation of Animate Objects and Actions. *Journal of Experimental Biology*, 146:87–113, 1989.

[PK02]   R. W. Picard and J. Klein. Computers That Recognise and Respond to User Emotion: Theoretical and Practical Implications. *Interacting With Computers*, 14:141–169, 2002.

[PMRR00]   J. P. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10), 2000.

[RP99]   M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–25, 1999.

[Ste04]   F. Stein. Efficient Computation of Optical Flow Using the Census Transform. In C. Rasmussen, H. Bülthoff, B. Schölkopf, and M. Giese, editors, *Pattern Recognition*, volume 3175 of *Lecture Notes in Computer Science*, pages 79–86. Springer Berlin / Heidelberg, 2004.

[SWP05]   T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *In CVPR*, pages 994–1000, 2005.

[TN11]   S. Tschechne and H. Neumann. Ordinal Depth from Occlusion Using Optical Flow: A Neural Model. In *Proceedings of Vision Sciene Society Meeting 2011 (VSS11)*, 2011.

[Tur04]   M. Turk. Computer vision in the interface. *Commun. ACM*, 47:60–67, January 2004.

[Ull84]   S. Ullman. Visual routines. *Cognition*, 18(1-3):97–159, 1984.

[UVNS02]   S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, 5(7):682–687, 2002.

[VPBP08]   A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social Signals, Their Function, and Automatic Analysis: A Survey. In *Proceedings of the 10th international conference on Multimodal interfaces*, ICMI '08, pages 61–68, New York, NY, USA, 2008. ACM.

[WADP97]   C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-Time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[WH99]   Y. Wu and T. S. Huang. Capturing Articulated Human Hand Motion: A Divide-and-Conquer Approach. *Computer Vision, IEEE International Conference on*, 1:606, 1999.

[WN08]   U. Weidenbacher and H. Neumann. Unsupervised Learning of Head Pose through Spike-Timing Dependent Plasticity. In *Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems*, PIT '08, pages 123–131, Berlin, Heidelberg, 2008. Springer-Verlag.

[WN09]   U. Weidenbacher and H. Neumann. Extraction of Surface-Related Features in a Recurrent Model of V1-V2 Interactions. *PLoS ONE*, 4(6), 2009.

[XP98]   C. Xu and J. L. Prince. Snakes, Shapes, and Gradient Vector Flow. *IEEE Transactions on on Image Processing*, 7(3):359–369, 1998.