MULTIMODAL–MULTISENSOR INTERFACES IN INDUSTRIAL AND MEDICAL APPLICATION DOMAINS
Multimodal-multisensor interfaces combine one or more user input modalities with sensor information (e.g., location, proximity, tilt). Sensor-based cues may be used to interpret a user’s physical state, health status, mental status, current context, engagement in activities, and many other types of information.

Sensor input aims to transparently facilitate user-system interaction, and adaptation to users’ needs. The type and number of sensors incorporated into multimodal interfaces has been expanding rapidly, resulting in explosive growth of multimodal-multisensor interfaces.

We believe that multimodal-multisensor interfaces can be designed to most effectively advance human performance during the next decade.

We demonstrate this by recent DFKI projects in industrial and medical application domains.
IUI AND ADVANCES IN INFORMATION EXTRACTION: FROM TEXT AND IMAGE TO KNOWLEDGE

- [http://www.dfki.de/~sonntag/courses/WS15/IUI.html](http://www.dfki.de/~sonntag/courses/WS15/IUI.html)
- [http://www.dfki.de/~sonntag/courses/SS14/IE.html](http://www.dfki.de/~sonntag/courses/SS14/IE.html)
Who was world champion in 1990?
Definition 5 Knowledge Base

A Knowledge Base (KB) is a structure

\[ KB := (C_{KB}, R_{KB}, I, \iota_C, \iota_R) \quad (2.12) \]

consisting of

- the sets \( C_{KB} \) and \( R_{KB} \) of supported concepts and relations within the knowledge base;
- the set \( I \) of instance identifiers (also called instances or objects);
- a function \( \iota_C : C_{KB} \rightarrow 2^I \) called concept instantiations;
- a function \( \iota_R : R_{KB} \rightarrow 2^{I^+} \) called relation instantiations,
\[ \forall r \in R : \iota_R(r) \subseteq \prod_{1 \leq i \leq |\sigma(r)|} \iota_C(\pi_i(\sigma(r))) \]
- a function \( \iota_A : R_{KB} \rightarrow I \times v_p(\pi_2(\sigma(a))) \) called attribute instantiation, with \( \iota_A(r) \subseteq \iota_C(\pi_1(\sigma(a))) \times v_p(\pi_2(\sigma(a))) \); \( v_p() \) being the values of the primitive datatypes \( p \in P \).
MULTIMODAL DIALOGUE SYSTEMS

ONTOGIES AND ADAPTIVITY IN DIALOGUE FOR QUESTION ANSWERING
**MULTIMODAL DIALOGUE SYSTEMS**

**INFORMATION STATE**

<table>
<thead>
<tr>
<th>Feature Class</th>
<th>IS State Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMR</td>
<td>Listening, Recording, Barge-in, Last-ok, Input dominance (text or voice)</td>
</tr>
<tr>
<td>NLU Query</td>
<td>Confidence, Domain relevance</td>
</tr>
<tr>
<td>NLU Fusion</td>
<td>Dialogue act, Question Foci, Complexity, Context object, Query text</td>
</tr>
<tr>
<td>NLU Answer</td>
<td>Fusion act, Co-reference resolution</td>
</tr>
<tr>
<td>NLU Manager</td>
<td>Success, Speed, Answer streams, Status, Answer type, Content, Answer text, Turn/Task numbers, Idle states, Waiting for Results, User/system turn, Elapsed times: input/output, Dialogue act history (system and user) e.g. reject, accept, clarify</td>
</tr>
</tbody>
</table>

![Diagram of Information State](image-url)
METACOGNITION

ANSWER STREAM PREDICTION
• Make dialogue-based radiology image reporting possible

• Reduce turn-over times and annotation errors

• Facilitate structured reporting

With the iPad’s FDA approval, a breakthrough for mobile medical imaging, especially in the U.S., can be expected. With RadSpeech, we aim to build the next generation of intelligent, scalable, and user-friendly mobile semantic search and image annotation interfaces for the medical imaging domain.
DIGITISATION AND DIGITALISATION

Knowledge Acquisition
Pi-Rads
DIGITAL PEN FEATURES

ONLINE FEATURES

\[ \bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mu_i \]

where \( n \) is the number of samples used for the classification, the mean radius \( \mu_i \) (standard deviation) as

\[ \mu_i = \frac{1}{n} \sum_{i=1}^{n} \| s_i - \bar{\mu} \| \]

and the angle \( \varphi_{i,i} \) as

\[ \varphi_{i,i} = \cos^{-1} \left( \frac{(s_i - s_{i-1}) \cdot (s_{i+1} - s_i)}{\| s_i - s_{i-1} \| \| s_{i+1} - s_i \|} \right) \]

A stroke is a sequence \( S \) of samples,

\[ S = \{ s_i | i \in [0, n-1], t_i < t_{i+1} \} \]

where \( n \) is the number of recorded samples. A sequence of strokes is indicated by

\[ D = \{ s_i | i \in [0, m-1], \} \]

where \( m \) is the number of strokes. The area \( A \) covered by the sequence of strokes \( D \) is defined as the area of the bounding box that results from a sequence of strokes.

<table>
<thead>
<tr>
<th>ID</th>
<th>Feature</th>
<th>Description</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Number of Strokes</td>
<td>( N )</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Length</td>
<td>( \lambda = \frac{\sum_{i=1}^{n} | \text{vecs}<em>{s_i} - \text{vecs}</em>{s_{i-1}} |}{n} )</td>
<td>( s ) denotes a sample.</td>
</tr>
<tr>
<td>2</td>
<td>Area</td>
<td>( A )</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Perimeter Length</td>
<td>( \lambda_p )</td>
<td>Length of the path around the convex hull.</td>
</tr>
<tr>
<td>4</td>
<td>Compactness</td>
<td>( c = \frac{\lambda_p^2}{A} )</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Eccentricity</td>
<td>( e = \sqrt{1 - \frac{b^2}{a^2}} )</td>
<td>( a ) and ( b ) denote the length of the major or minor axis of the convex hull, respectively.</td>
</tr>
<tr>
<td>6</td>
<td>Principal Axes</td>
<td>( e_i = \frac{b_i}{a_i} )</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Circular Variance</td>
<td>( \nu_c = \frac{1}{n \mu_c^2} \sum_{i=1}^{n} (| \text{vecs}_{s_i} - \bar{\mu} | - \mu_c)^2 )</td>
<td>( \mu_c ) denotes the mean distance of the samples to the centroid ( \mu_c ).</td>
</tr>
<tr>
<td>8</td>
<td>Rectangularity</td>
<td>( r = \frac{A}{ab} )</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Closure</td>
<td>( c_i = \frac{| s_i - s_{i+1} |}{\lambda} )</td>
<td>( \varphi_i ) denotes the angle between the ( s_k ), at ( s_i ) and ( s_{i+1} ) at ( s_i ).</td>
</tr>
<tr>
<td>10</td>
<td>Curvature</td>
<td>( k_i = \frac{2}{| s_i - s_{i+1} |} )</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Perpendicularity</td>
<td>( P_i = \frac{\sum_{i=1}^{n} \sin(\varphi_i)}{n} )</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Signed Perpendicularity</td>
<td>( P_i = \frac{\sum_{i=1}^{n} \sin(\varphi_i)}{n} )</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Angles after Equidistant Resampling (6 line segments)</td>
<td>( \sin(\alpha) ), ( \cos(\alpha) )</td>
<td>The five angles between succeeding lines are considered to make the features scale and rotation invariant (normalization of writing speed).</td>
</tr>
</tbody>
</table>

**Gestures:**
- Circle
- Line
- N
- Box
- Checkmark
- Arrow

**Interpretations:**
- Free Text Area: character “O/O”, or “0” according to the text field grammar
- Sketch Area: position of specific area or coordinate
- Annotation Vocabulary Fields: marking of a medical ontology term
DIGITAL ENHANCEMENT / BLENDED INTERACTION / TANGIBLE INTERACTION

EXAMPLES OF BLENDED INTERACTION RANGE FROM DIGITAL PEN & PAPER AND PEN & MULTI-TOUCH INTERACTION TO TANGIBLE DISPLAYS IN ROOMS OF MIXED OR AUGMENTED REALITY.

Combines the virtues of physical and digital artifacts.

The quality of Blended Interaction is judged by its compatibility with our natural cognitive processes when we interact and collaborate in the real non-digital world.

<table>
<thead>
<tr>
<th>System Features</th>
<th>Paper</th>
<th>Mammo Digital Paper</th>
<th>PC (iSoft)</th>
<th>ASR (Nuance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pen-on-paper interface</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Immediate validations</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Offline validation (of digital content)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Realtime recognition (text)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Realtime recognition (gestures)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online correction of recognition errors</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real-time capture to structured database</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forward capture to database</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Source Document (Certificate)</td>
<td>x</td>
<td></td>
<td>(x)</td>
<td></td>
</tr>
<tr>
<td>Digital Source Document (Certificate)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training hours before effective usage</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>No user distraction from primary task</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No distraction from patient</td>
<td>x</td>
<td>x</td>
<td>(x)</td>
<td></td>
</tr>
<tr>
<td>Average time to complete one predefined Radlex entry</td>
<td>3 sec</td>
<td>3 sec</td>
<td>5 sec</td>
<td>2 sec</td>
</tr>
</tbody>
</table>

The quality of Blended Interaction is judged by its compatibility with our natural cognitive processes when we interact and collaborate in the real non-digital world.

The chart shows the time for structured reporting for different methods: Paper, Digital Pen, PC, and Speech.
Digital manufacturing
COMPLEX MULTIMODAL MULTISENSOR SYSTEMS

ERmed
Erweiterte Realität in der Medizin

Not Looking at Screen

XML-RPC
ACTIVITY RECOGNITION

Companion technology and social interaction
Episodic memory event encoding model (Breakfast Scenario)

Sensor Data

- Eye tracker (with scene camera)
- GPS, or other sensors

Interpretation of raw sensor data: e.g., object recognition, location estimation, ...

Attention to ...
- Face
- Object
- Text

Encoded Event

- Location: Living room
- Activity: Make a sandwich
- Person A: Speak with Person A
- Texture: "take 1 pill in the morning..."
- Activity database: (created by crowdsourcing platform: LabelMovie)
- Object database (including faces): "ingredients:..."

Activity database

Object database

GPS, or other sensors

Information processing

- Interpreting sensor data
- Encoding observations
- Creating episodic events

Object recognition
Location estimation

Databases and recognition modules

Object DB

- id: "bread", type: "object", image: ["sample1.png", ...], features: ["feature1.txt", ...], description: "bread is a food"

Activity DB

- id: [UNIQUE ID], start: "2014/10/30/10:05:54", end: "2014/10/30/10:20:12", activity: "discuss", object: "Takumi"

Episodic Memory DB

- id: "eat", level: 2, derived_from: ["bite", "chew", ...], form: "have_a_meal"
- id: "have_a_meal", level: 3, derived_from: ["eat", "drink", ...], form: ""
- id: "discuss", level: 2, derived_from: ["look_at_face", "speak", ...], form: "meeting"

Face Recognizer

- Person A

Object Recognizer

- Bread

Text Recognizer

- "aspirin"

Kognit Cloudant Database: https://kognit-tt.cloudant.com/
• By combining image analysis technologies with eye gaze analysis technologies, a computer can recognise the visual content the user is attending to.

• Analysis of user eye gaze is useful to present information of attended content in an adequate way.

• Gaze Videos (x2)

Real-Time Gaze-based Object Recognition
TECHNOLOGY AND PROJECTS

CONNECTING THE DOTS WITH PROJECTS AND SYMPOSIA

SMARTWEB
THESEUS
RADSPEECH
MEDICAL CPS
ERMED
AAAI
AAAI
AAAI
SMART FACTORIES
KOGNIT
KDI
MOBILE REFUGEE
INTERA-KT
PERSUASIVE

timeline
2004
2017
2018
2019
DEEP LEARNING BASED EPISODIC MEMORY

Object Recognition
20 -> 2000 objects
Head-worn Eye Tracker

(a) Task completion time
(b) Number of utterances
DEEP LEARNING FOR OBJECT DETECTION

<table>
<thead>
<tr>
<th>Input</th>
<th>Dataset</th>
<th>Model</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>iconic</td>
<td>ImageNet</td>
<td>BVLC GoogLeNet</td>
<td>Caffe</td>
</tr>
<tr>
<td>non-iconic</td>
<td>ILSVRC-2012</td>
<td>R-CNN</td>
<td>Faster R-CNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YOLO</td>
<td>ELTE R-CNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>YOLO</td>
</tr>
</tbody>
</table>
NEED FOR APPLICATION SPECIFIC DATA

MULTIMODAL MULTISENSOR ACTIVITY ANNOTATION TOOL
WHY MULTIMODAL-MULTISENSOR INTERFACES HAVE BECOME DOMINANT

- Flexibility

- They support users’ ability to select a suitable input mode, or to shift among modalities as needed, during the changing physical contexts and demands of continuous mobile use.

- Likewise there are ideal for supporting individual differences among users.

- This has stimulated the paradigm shift toward multimodal-multisensor interfaces on computers today, which often is further enhanced by either multimodal output or multimedia output.
2 PART OF ABSTRACT

- Do not think about the best method of machine learning—the central issue is the application domain and domain problems such as integrated decision support for doctors.

- More of IUI’s effort should go into real-life problems, approaches, and architectures.

- Our cognitive computing approach is an incremental knowledge acquisition process rather than almost exclusively using a data-driven engineering model of research.
Incremental Knowledge Acquisition

1. Structured/Structural Knowledge
   - Representational Ontology: RDFS, OWL
   - Upper Ontology: time, space, organization, person, event
   - Information Element: Ontology images, texts, vocabularies
   - Clinical Ontology: doctor, nurse, patient, medical case - DICOM Ontology
   - Thesauri & Taxonomies

2. Desktop Annotation
   - "Search comparable Lymph cases"

3. Automatic Image recognition

4. Spatial reasoning

Dicom
SENSOR-ACTION-LOOP IN CPS (IN KOGNIT)
COGNITIVE ENHANCEMENT

improving cognitive abilities
Towards cognitive enhancement

Compensate for the physical and sensory deficits that may accompany aging

- lift chair, wheel chair
- ergonomic handles

no computer technology

Advanced computer-based technologies for AAL (ambient assisted living)

- SSPI
- exoskeleton
- control household appliances (using, e.g., head gestures)

- hearing aid device
- cardiac pacemaker

computer technology

Assurance of, compensation for, assessment of cognitive deficits
CIND / Dementia

AI technology

- SSPI (Speech)
- AI companion

sensor-motor and psychosocial issues

cognitive decline

Compensation Paradox

- User must be made aware of planned task/activity and must be guided.
- User and caregiver satisfaction - usability / utility.
- Avoid introducing inefficiency into user activities - usability / utility.
- Avoid making the user overly reliant on the compensation system?
- Request confirmation about whether an activity has been completed successfully.
Kognit Hardware Overview

- Narrative Clip
- Pupil Labs Eye-tracker
- SMI Eye-tracker
- Samsung Galaxy Note 4
- Cybershot scene cam
- Anoto
- WheelPhone
- NAO Humanoid
- Leap Motion
- Accu LED projector
- Low range 3D cam
- Structure Sensor
- Tobii EyeX
- Brother Airscouter
- Oculus DK 2
- Space Glasses Meta Pro (3D cam)
- Epson Moverio BT-200
Gehirnersatz für Demente

Daneil Sperig Experte für künstliche Intelligenz


Nimm jetzt Deine Tabletten

10:30
CONCLUSION, LESSONS LEARNED AND FUTURE INVESTIGATIONS
CONCLUSIONS

ML IS NOT A SILVER BULLET

- The interest has also being fueled by the recent research breakthroughs brought about by deep learning.
- Currently ML has several limitations in complex real-life situations. Some of these limitations include:
  - many ML algorithms require large number of training data that are often too expensive to obtain in real-life,
  - significant effort is often required to do feature engineering to achieve high performance,
  - many ML methods are limited in their ability to exploit background knowledge,
  - lack of a seamless way to integrate and use heterogeneous data.
  - do not provide end-to-end intelligent user interface systems.
CONCLUSION

DEEP LEARNING

- Success stories: supervised learning, backpropagation, stochastic gradient decent, massive amounts of data. Convolutional NN, RNN.

- Deep Learning in practice can be a game changer (95%->99%).
  - Parallel is good, layering is good.
  - Important Tool box

- But: technology can be replicated, data cannot.
  - data acquisition (digitisation and digitalisation) is a strategic goal
  - what’s your multimodal-multisensor data strategy over the long run?
Collect annotated images (segmentation / bounding boxes) from relevant application domains.

- Annotate images from, e.g., Google Images
- Annotate images from own application domain

- Increase dataset
- Rotate
- Set background
- Set objects
- Occlude with additional objects (if possible from application domain)
- Add surrogate semantics to Faster R-CNN and R-FCN

Add models and semantics:
- Head / hand model
- ConceptNet

Interactive machine learning,
iteratively add training examples.
Tight interaction loop.
Computer is part of a human design process.
Interactively edit data.

How does a human-centred approach change the way machine learning is done?
FUTURE INVESTIGATIONS

- rule-based systems, ontologies and deep learning
- self-training in narrower contexts
- anomaly detection and single-shot-learning
- more sensors
- goal-orientation
- episodic memory in VR
- providing examples of cognitive computing