Dokumentation
Demonstrator SmartWeb Erlangen
"Klassifikation des Benutzerfokus"

Christian Hacker
Lehrstuhl für Mustererkennung (LME)
Universität Erlangen-Nürnberg
hacker@informatik.uni-erlangen.de

27. März 2007

Inhaltsverzeichnis

1 Grundlegendes

Demonstriert wird die Klassifikation des Benutzerfokus (On-Focus/Off-Focus) aus dem Video- und Audio-Signal (Abb. 1). Im Audio-Signal wird On-Talk von verschiedenen Off-Talk Klassen unterschieden, im Videosignal die Blickrichtung (On-View vs. Off-View). Möglichkeiten, die im SmartWeb Szenario auftreten können, sind in Tab. 1 aufgeführt.

Die On-/Off-View Klassifikation erfolgt bildweise mit dem Viola-Jones Algorithmus. On-/Off-Talk wird wortweise mit prosodischen Merkmalen und neuronalen Netzen berechnet. Die satzweise Klassifikation aus Video und Audio sowie die Fusion erfolgt wahlweise durch Mittelung der wort- und bildweisen Ergebnisse (und Schwellwerten zur Fusion) oder standardmäßig durch drei weitere Klassifikatoren.

Abbildung 1: Klassifikation des Benutzerfokus.

	On-View	Off-View
NOT (On-Talk)	**On-Fokus, Interaktion mit dem System**	*(kann evtl. vorkommen)*
ROT	Lesen vom Display	—
POT	*(kann evtl. vorkommen)*	Über die Auskunft von SmartWeb berichten
SOT	Auf eine Unterbrechung reagieren	Auf eine Unterbrechung reagieren

Tabelle 1: On-Fokus vs. Off-Fokus. Audio: Gelesener Off-Talk (**ROT**), Paraphrasing Off-Talk (**POT**), Spontaneous Off-Talk (**SOT**), No Off-Talk (**NOT**). Video: On-View/Off-View

2 Vorbereitung

Voraussetzung ist ein Linux-System. Getestet wurde mit SuSE 10.0 und SuSE 10.1.

2.1 Kamera

Getestet wurde das Modul mit einer Firewire-Kamera (*Unibrain fire-i*) und einigen USB-Kameras (z.B. *Logitech QuickCam Pro 4000*). Für die USB-Kameras ist der pwc-Treiber empfehlenswert. Die Kamera *Logitech Quickcam Messenger* funktioniert mit dem gspca-Treiber.

Kameras die mit dem pwc-Treiber funktionieren sind unter http://www.lavrsen.dk/twiki/bin/view/PWC/WorkingWebcamsWithPWC aufgelistet. Eine Negativliste ist unter http://www.saillard.org/linux/pwc/ zu finden. Dort kann man auch die Treiber-Module downloaden.

Der pwc-Treiber ist ganz einfach zu installieren:

- Download `pwc-10.0.12-rc1.tar.bz2` von http://www.saillard.org/linux/pwc/files
- `bunzip2 pwc-10.0.12-rc1.tar.bz2; tar xf pwc-10.0.12-rc1.tar`
- `make`
- Als ROOT: `modprobe -r pwc`
- Als ROOT:
  `cp pwc.ko /lib/modules/2.6.16.27-0.6-smp/kernel/drivers/usb/media/pwc/` wobei 2.6.16.27-0.6-smp die Kernel-Version ist (`uname -r`)
- Als ROOT: `depmod -a`
- `modprobe pwc` (Nach jedem Neustart, ggf. für den Benutzer mit `visudo` erlauben)

Testen kann man die Kamera z.B. mit `gnomemeeting` oder

```
bin_Linux/onview -in usb /dev/video0 -cammode 5 -timeout -1 -disp -cascade
config/FaceClassifier_M06.xml
```

Achtung: Neustart bei eingesteckter Kamera führt u.U. dazu, dass die Kamera als Standard-Soundkarte verwendet wird.

"Version without firewire cam support" unter
`bin_Linux/onview -h`
gibt an, dass keine Firewire Module aktiviert/installiert sein müssen (video1394, raw1394).

2.2 Mikrophon

Es ist wichtig, dass das Mikrophon **ähnlich ausgesteuert ist, wie bei der Aufzeichnung der Trainingsdaten**. Als gut erwies sich , den Eingangsregler auf 1/3 zu stellen und Mic Boost (+20 dB) einzuschalten. Ein Beispiel für On-Talk ist

```
DATA/BEISPIELE/BeispielOntalk.ssg
```

`*.ssg` bezeichnet Audio-Dateien ohne Header (raw). Diese sind abspielbar mit:

```
sox -r 8000 -t sw DATA/BEISPIELE/BeispielOntalk.ssg -t ossdsp /dev/dsp
```

Eine genaue Aussteuerung der Aufnahme ist auch noch in der GUI möglich.

Achtung: Neustart bei eingesteckter Kamera führt u.U. dazu, dass die Kamera als Standard-Soundkarte verwendet wird.

2.3 Quicktime

`bin_Linux/onview -h`
"Version without quicktime movie support", gibt an, dass die Quicktime4Linux Bibliothek nicht installiert sein muss. Allerdings lassen sich dann auch keine Videos klassifizieren/aufzeichnen.

2.4 Benötigte Programme

Er werden Standardprogramme benötigt wie

`sox`, `dd`, `grep`, `sort`, `tail` oder `cat`,

da die GUI in einer Skript-Sprache realisiert ist. Von der GUI werden die SmartWeb-Module `ontalk` und `onview` sowie der LME-Spracherkenner aufgerufen und die Ergebnisse geparst.

Ferner muss perl und perl Tk installiert sein und Tk.pm im Pfad PERL5LIB zu finden.

3 Der Demonstrator

3.1 Start

die Verzeichnisse
`onview/`
`sympalog/`
und `tmp/`

befinden sich parallel in `$TESTBED_HOME`. Der Demonstrator liegt in `onview/`, alle temporären Dateien sowie Aufnahmen werden nach `tmp/` geschrieben. Wird des SmartWeb-Erkenner der Firma Sympalog in der GUI optional ausgewählt, so wird auch `sympalog/` benötigt.
Gestartet wird der Demonstrator mir

```
tcsh
setenv TESTBED_HOME /home/hacker/smartweb
setenv TESTBED_TMP $TESTBED_HOME/tmp
DemoOnfocus 10.0
DemoOnfocus 10.1
```

je nach SuSE-Version (ab 10.1 verlaufen ohne diese Unterscheidung die "Progressbars" leider von oben nach unten). Default ist 10.1. Es öffnet sich die GUI und ein separates Fenster mit dem Kamerabild (Abb. 3.3). Dieses läßt sich vergrößern/verkleinern (Tasten +/- oder mit der Maus). "Fotos knipsen" kann man mit Taste c.

Fehlerbehandlung: Sollte sich das Kamera-Fenster nicht öffnen, kann das folgende Ursachen haben:

- `/dev/video0` oder `/dev/video1` ... existieren nicht, oder Benutzer ist nicht in der richtigen Gruppe.
  `ls -s /dev/video0`
  `crw-rw---- 1 root video 81, 0 Feb 19 09:45 /dev/video0`
- Kameramodule sind nicht geladen: `modprobe pwc` (siehe Abschnitt 2.1).
- `bin_Linux/onview` läuft bereits. Abhilfe: `bin_Linux//killonview.pl`

Andere Aufrufmöglichleiten:
Demonstration ohne Video:
`DemoOntalk 10.1`

3.2 Ablauf einer Vorführung

1. `setenv TESTBED_HOME /home/hacker/smartweb/`
   `setenv TESTBED_TMP $TESTBED_HOME/tmp`
2. `./DemoOnfocus 10.`x, $x \in \{0, 1\}$
3. Datei zum Schreiben öffnen:
   [New...] → test1 → [save] → [yes]
4. Aufnahme:
   [REC] → *"Gibt es ein italienisches Restaurant in der Nähe"* → [STOP]
   Die Auswertung des Videosignals wird mit den Balken On-View und Off-View visualisiert.
5. Abspielen: [PLAY]
6. Nur bei der ersten Aufnahme pro Session: Aussteuern des **On-Talk** Signals in der Menüleiste:
   [Micro] → [Adjust Volume]
7. Spracherkennung: [Speech Recognition]
   Im rechten Feld erscheint die erkannte Wortfolge und die Zeitzuordnung.
8. Klassifikation mit Hilfe der Zeitzuordnung aus dem letzten Schritt:
   [Classification of On-/Off-Talk]
   - Im rechten Feld erscheint die wortweise Klassifikation des Audio-Signals zusammen mit den wortweisen Wahrscheinlichkeiten.
   - Die mittleren Balken On-Talk und Off-Talk zeigen (satzweise) das Ergebnis aus dem Audiosignal an.
   - Die rechten beiden Balken On-Focus und Off-Focus zeigen das Ergebnis der Fusion an.
9. Nun kann erneut mit Punkt 3 fortgefahren werden (ohne Punkt 5).

3.3 Tipps für eine gelungene Vorführung

- Die Lautstärke muss richtig ausgesteuert sein, siehe Abschnitt 2.2 und die Anweisungen in Punkt 5 in Abschnitt 3.2. Wichtig ist, dass [Adjust Volume] für einen **On-Talk**-Turn durchgeführt wird.
- Bespielsätze (weitere in Tab. 2 oder im Abschnitt 3.4.6)
  On-Talk: *"Gibt es ein griechisches Restaurant in der Nähe"*
  Off-Talk: *"Warte mal kurz"*
  Off-Talk: *"Dort drüben müsste das Restaurant sein"*
  Off-Talk: *"So würde es klingen, wenn ich zu jemandem anderen spreche"* (nicht im Vokabular)
  gelesener Off-Talk (ROT): *"Goldener Hecht – zur Post – Akropolis"* (fiktives Feedback am Display des Systems)
- On-Talk sollte laut und deutlich gesprochen werden; Tonhöhe variieren!

On-Talk:	SmartWeb, gibt es hier ein indisches Restaurant?
On-Talk:	Ich möchte was schnelles essen.
On-Talk:	Gibt es eine Fußball Kneipe in der Nähe?
On-Talk:	Gibt es ein Cafe in der Nähe?
On-Talk:	Gibt es in der Nähe ein Restaurant?
On-Talk:	Wo ist der nächste Schnellimbiss?
On-Talk:	Gibt es einen Stadtplan mit dem Cafe?
Gelesener Off-Talk:	Kreta , Max-Planck-Straße vierundzwanzig
Gelesener Off-Talk:	Brasserie , Kulisse , Pizzeria-Bruno , Burger-King
Gelesener Off-Talk:	Poseidon , Max-Planck-Straße vierundzwanzig
Off-Talk:	Gleich in der Nähe ist ein spanisches Restaurant!
Off-Talk:	Gleich dort drüben müsste ein Cafe sein.
Off-Talk:	Ich frag mal nach.
Off-Talk:	Waret kurz!

Tabelle 2: Beispiele im Sprachmodell Restaurant

- Off-Talk sollte etwas leiser gesprochen werden (der Nutzer will das System nicht verwirren) und etwas undeutlicher, spontaner, vielleicht auch schneller. Eintönig, rauh, unmelodisch.
- Gelesener Off-Talk ist leiser und nachdenklich langsam.
- Die Anweisungen für den LME-Schauspiel-Korpus waren: Off-Talk leiser, gelesener Off-Talk zusätzlich langsamer. So wurden allein vor der Fusion über 90% Erkennungsrate erzielt.
- Für die spontanen SmartWeb Daten der LMU München wurden keine Anweisungen gegeben. Die Erkennungsrate ist über 80 %, jedoch erst nach der multimodalen Fusion.

3.4 Detaillierte Beschreibung

3.4.1 Aufnahme [New...], [REC]

Datei zum Schreiben öffnen: [New...] → test1 → [save] → [yes].

Die Aufnahme erfolgt mit der [REC]-Taste, die sich dann zu [STOP] verändert. [REC] → *"Gibt es ein italienisches Restaurant in der Nähe"* → [STOP] Die Auswertung des Videosignals wird mit den Balken On-View und Off-View visualisiert.

3.4.2 Beispiel-Dateien öffnen: [Open...]

Statt mit [New...] eine schreibbare Datei zu öffnen, können mit [Open...] Dateien geladen werden, z.B. `DATA/BEISPIELE/BeispielOntalk.ssg`. Es wird nur das Audio-Signal ausgewertet.

Wenn das `onview`-Modul mit Quicktime-Support vorliegt (siehe Abschnitt 2.3), können unter `DATA/MULTIMEDIA/ssg/` auch Dateien geöffnet werden, für die zusätzlich die Videodatei in `DATA/MULTIMEDIA/mov/` ausgewertet wird.

3.4.3 Aufnahmepegel: [Micro]

Im Menü "Micro" kann man mit [Adjust Volume] den Aufnahmepegel für ein On-Talk-Signal kalibrieren, falls die Aufnahme mit dem Mixer nicht optimal eingestellt ist (siehe Abschnitt 2.2). Der aktuelle Faktor mit dem das Sprachsignal multipliziert wird, ist in [Set Volume] einzusehen und kann manuell angepasst werden.

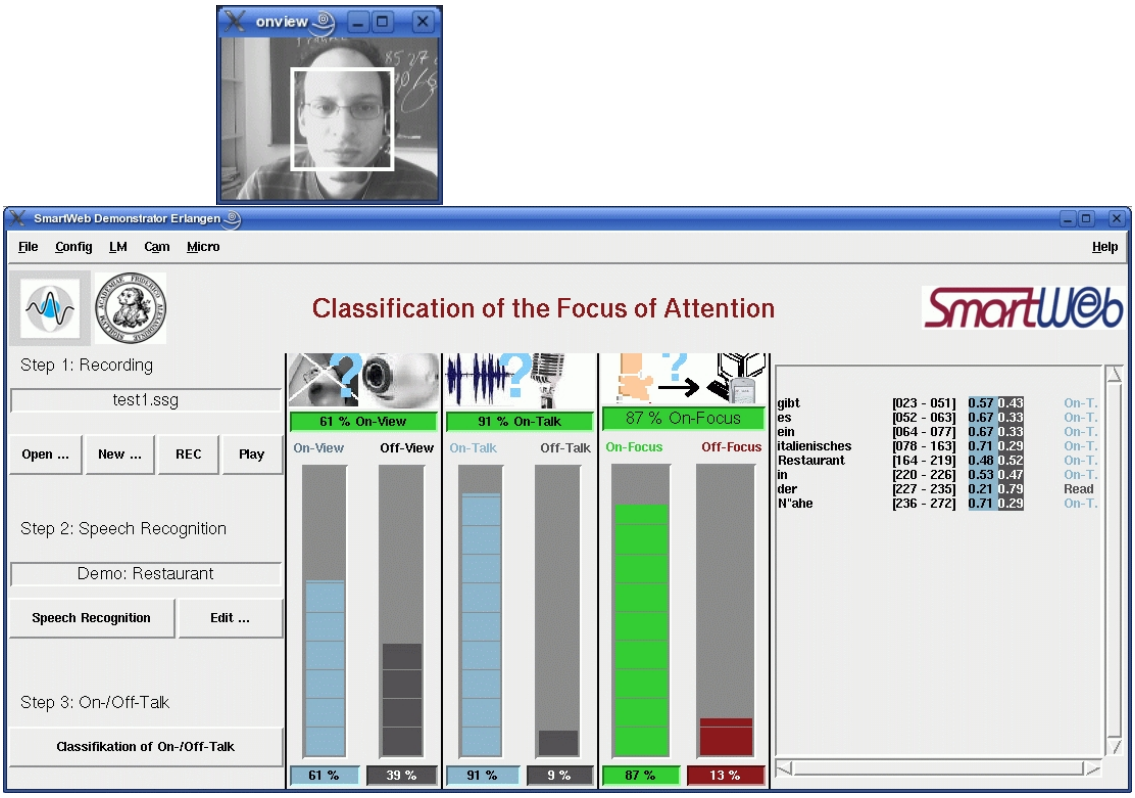

Abbildung 2: Der Demonstrator.

3.4.4 Kamera starten: [Cam]

Im Menü "Cam" lässt sich On-View/Off-View starten und stoppen. (Stoppt automatisch, wenn mit [Open...] eine Multimedia-Datei geöffnet wird.)

3.4.5 Das Kamera-Fenster

Dieses QT-Widget läßt sich mit der Maus oder mit +/- Tasten vergrößern/verkleinern. "Fotos knipsen' mit Taste c;

3.4.6 Sprachmodelle und Beispielsätze [LM]

Im Menü "LM" können verschiedene einfache Sprachmodelle ausgewählt werden, die bei der Demonstration eine robuste Spracherkennung gewährleisten. Es kann zwischen "Restaurant", "Fußball", "Nahverkehr" und einer Kombination ausgewählt werden. Dieser Spracherkenner ist *nicht* der Erkenner, der im SmartWeb Gesamtsystem integriert ist. Der integrierte LME-Erkenner wurde auf szenario-fremden SmartKom-Daten trainiert; die Akkuratheit ist nicht optimal, wird aber durch ein einfaches Sprachmodell verbesssert. Beispielsätze des jeweils gewählten Sprachmodells werden zufällig unter "LM" → [Examples...] erzeugt. Einige Beispiele sind in den Tabellen Tab. 2 (Restaurant), Tab 3 (Nahverkehr) und Tab 4 (Fußball) zu finden.

Das Sprachmodell "SymRecSW" verändert nicht nur den Wortschatz auf den aktuellen SmartWeb Gesamtwortschatz, sondern aktiviert auch den Sympalog SmartWeb-Erkenner samt OOV-Erkennung. Dazu ist das SmartWeb-Verzeichnis `$TESTBED_HOME/sympalog/` erforderlich. Die OOV-Erkennung wird künftig in der GUI noch datailierter dargestellt.

On-Talk:	Wann fährt hier die nächste U-Bahn?
On-Talk:	Gibt es einen Stadtplan mit der Haltestelle?
On-Talk:	Wie lange fährt man mit der U-Bahn zum Hauptbahnhof?
On-Talk:	Wann fährt hier die nächste S-Bahn?
On-Talk:	Bis welche Uhrzeit fährt in der Nacht die S-Bahn?
On-Talk:	Wie teuer ist der Bus?
On-Talk:	Wie lange läuft man zum Messegelände?
On-Talk:	Gibt es eine Familienkarte?
Gelesener Off-Talk:	Linie neun bis vier Uhr , Linie sechs bis zwei Uhr dreißig
Gelesener Off-Talk:	Linie zwei , acht Stationen , dann in die eins umsteigen
Gelesener Off-Talk:	Vierzehn Uhr fünf , vierzehn Uhr elf , vierzehn Uhr sechsundzwanzig
Off-Talk:	Es gibt Busse bis vier Uhr.
Off-Talk:	Die S-Bahn fährt bis zwei.
Off-Talk:	Wir müssen mit der U-Bahn fahren.
Off-Talk:	Das sind über zwei Kilometer.
Off-Talk:	Wartet mal kurz.

Tabelle 3: Beispiele im Sprachmodell Nahverkehr

3.4.7 Kanaladaption des Spracherkenners [Config]

Die Werte zur adaptiven Kanaladaption sollten jeweils beim ersten Einsatz mit einem neuen Mikrophon zurückgesetzt werden: Menü "Config" → "Configure On-Talk" → [Turnbewertung] → [Initiale Adaptionsparam / mean.start] → [OK]

3.4.8 Spracherkennung [Speech Recognition]

Der aktuelle ausgewählte oder aufgezeichnete Satz wird erkannt und im rechten Fenster dargestellt. Ist das Sprachmodell "SymRecSW" aktiviert und wird somit der Sympalog SmartWeb-Erkenner verwendet, so wird beim ersten Aufruf der komplette Sympalog-Server gestartet (dauert einige Sekunden).

3.4.9 Manuelle Anpassung der Spracherkennung [Edit...]

Mit [Edit..] kann die erkannte Sprachkette manuell angepasst werden (erzwungene Zeitzuordnung).

Ist das Sprachmodell "SymRecSW" aktiviert und wird somit der Sympalog SmartWeb-Erkenner verwendet, so ist diese Funktion nicht verfügbar.

3.4.10 Graphiken [Help]

Im Menü "Help" befinden sich Graphiken zu SmartWeb, On-Talk/On-View, On-Talk-Merkmalen, On-View-Merkmalen und zur Fusion.

3.4.11 Multimodale Fusion [Config]

Zur Fusion werden Meta-Merkmale (Menue "Help" → [Fusion]) oder heuristische Schwellwerte verwendet. Die Standardeinstellung im Menü "Config" ist "On-Talk with Meta-features".

Literatur

[1] Moritz Kaiser, Hannes Mögele, Christian Hacker, and Anton Batliner. Multi-Modal Focus of Attention: Elicitation, Annotation, and Classification. *submitted to: Language Recources and Evaluation*, 2007.

On-Talk:	Welche Mannschaften spielen in Berlin?
On-Talk:	Welche Mannschaften spielen in Köln?
On-Talk:	Welche Spielorte gibt es?
On-Talk:	Wann spielt Tschechien?
On-Talk:	Welche Mannschaften spielen am Dienstag in München?
On-Talk:	Wann ist das nächste Spiel von Argentinien?
On-Talk:	Welche Spielorte gibt es bei der Fußball-WM?
On-Talk:	Wieviele Tore hat Spanien geschossen?
On-Talk:	Gegen wen spielt heute Deutschland?
Gelesener Off-Talk:	USA Frankreich eins zu null , Spanien USA zwei zu eins , USA Holland eins zu zwei
Gelesener Off-Talk:	elfmeter.de , chat.de, Bundesliga.de
Gelesener Off-Talk:	Niederlande gegen Brasilien , Deutschland gegen Spanien , Brasilien gegen England
Off-Talk:	England spielt heute gegen Frankreich.
Off-Talk:	Heute spielt Polen , das muss ich mir ansehen!
Off-Talk:	Brasilien hat gegen Argentinien eins zu zwei verloren.
Off-Talk:	Mist , es fängt an zu regnen.
Off-Talk:	Wartet kurz.

Tabelle 4: Beispiele im Sprachmodell Fußball

[2] Christian Hacker, Anton Batliner, and Elmar Nöth. Are You Looking at Me, are You Talking with Me – Multimodal Classification of the Focus of Attention. In P. Sojka, I. Kopecek, and K. Pala, editors, *Proc. Text, Speech and Dialogue. 9th International Conference, TSD 2006* , Lecture Notes in Artificial Intelligence , pages 581 – 588, Berlin, Heidelberg, 2006. Springer.

[3] A. Batliner, C. Hacker, and E. Nöth. To Talk or not to Talk with a Computer: On-Talk vs. Off-Talk . In K. Fischer, editor, *How People Talk to Computers, Robots, and Other Artificial Communication Partners* , University of Bremen, SFB/TR 8 Report , pages 79–100, 2006.

[4] Anton Batliner, Viktor Zeissler, Elmar Nöth, and Heinrich Niemann. Prosodic Classification of Offtalk: First Experiments. In *Proc. of the Fifth International Conference on Text, Speech, Dialogue*, pages 357–364, Berlin, 2002. Springer.