Vortrag im Seminar „Humanoide Roboter“ WS 2004/2005

KISMET

Patrick Artner

28. Februar 2005

Inhaltsverzeichnis

1 Motivation: Wozu dient Kismet

Entwickelt wird Kismet am MIT in der *Humanoid Robotics Group* im *Sociable Machines Project* zur Modellierung von kindlichem Lernen und Interaktion mit Menschen. Dr. Cynthia Breazeal ist seit 1998 am Projekt maßgeblich beteiligt (auch ihre Doktorarbeit handelt von Kismet).

Das Projekt hat viele (wie in USA üblich: militärische) Sponsoren:

- Office of Naval Research (ONR) - USA
- Defense Advanced Research Projects Agency (DARPA) - USA
- NTT (Nippon Telephon and Telegraphie Group) - Japan

Ziel ist die erschaffen eines *intuitiven* Interfaces für autonome Roboter deren Einsatz zum Beispiel im Bereich Haushalt, Unterhaltung oder Gesundheitspflege denkbar ist. Es wird versucht die Interaktion intuitiv, d.h. ohne großen Lernaufwand für den Nutzer, zu gestalten. Dabei wird ausgenutzt, das Menschen sehr gut im deuten von Mimik/Gestik anderer Menschen sind.
Kismet imitiert ein sehr junges Kind und animiert andere sich um ihn zu kümmern und versucht dann zu interagieren.

Der Benutzer kann aus Kismets Reaktionen/Gesichtsausdrücken Rückschlüsse auf die Auswirkung seines Verhaltens auf Kismet gewinnen. Erreicht wird dies dadurch daß Aussehen und Verhalten von Kismet einem jungen Kind nachempfunden sind. Aktionen und Reaktionen sind auch kindlich: sehr deutliche Mimik, Brabbelsprache.

1.1 Kismets Evolution anhand von Bildern

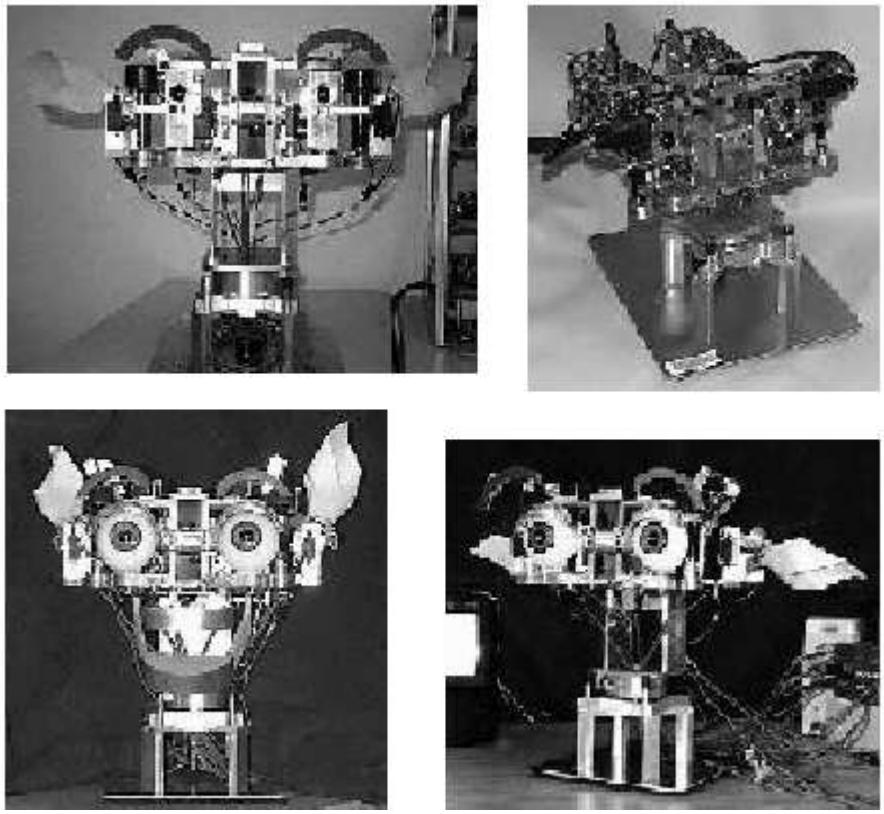

Abbildung 1: ältere Kismets (vor 1999)

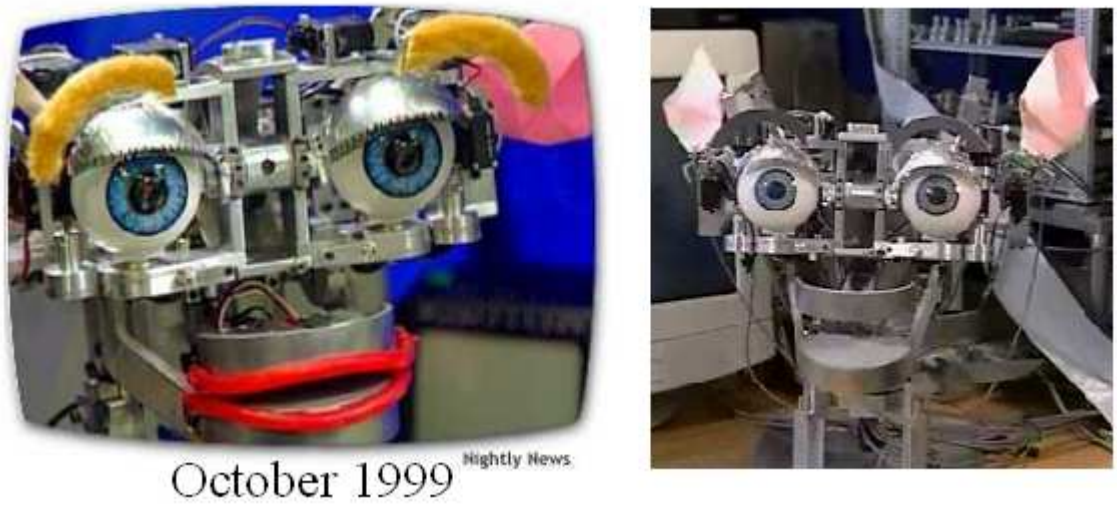

Abbildung 2: Kismets in der Presse, Kismet aus Video

2 Hardware

2.1 Roboterkopf

2.1.1 Motorik

Kismet ist ein menschenähnlicher Roboterkopf. Er besitzt gesamt 21 Degree of Freedom (DOF), davon 15 für den Gesichtsausdruck, 3 für die Augenbewegung und 3 um den Kopf selbst zu bewegen. Kismet nutzt seine Gesichtsausdrücke um seinen inneren Zustand darzustellen. Die Gesichtsausdrücke sind menschenähnlich modelliert und ermöglichem dem Interakteur eine intuitive Einsicht in Kismets „Gefühlswelt“.
Dies soll die Interaktion zwischen Kismet und Menschen erleichtern.

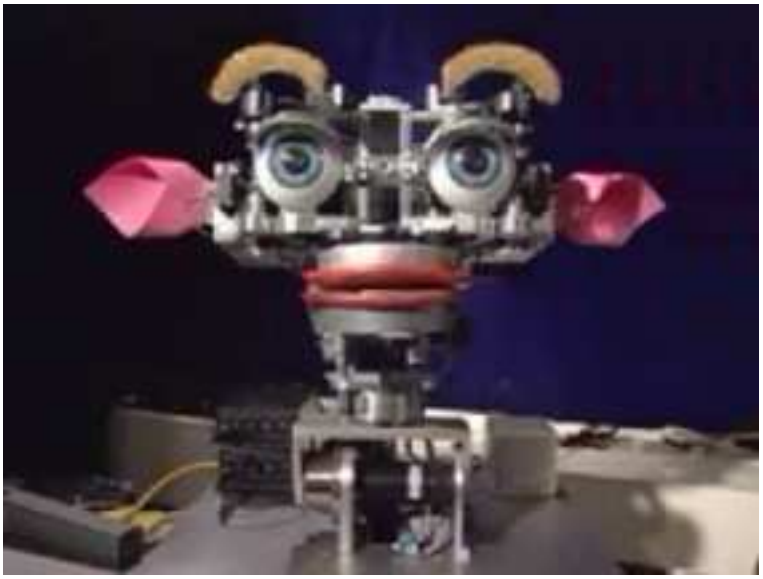

Abbildung 3: Kismet mit neutralem Gesichtsausdruck

Gesichtsausdruck			
Anzahl	DOF	Ort	Aktionen
2	2	Ohr	anheben/senken, nach vorne/hinten rotieren
2	2	Augenbraue	anheben/senken, Nasenwärts hin/weg verschieben
2	2	Lippe	Munwinkel anheben/senken
2	1	Augenlid	Augen auf/zu
1	1	Kinn	Mundbewegungen (Sprachsimulation)
	15	Simulation von Gesichtsausdrücken um Menschen einen intuitiven Ansatz zur Interaktion mit Kismet zu ermöglichen	

Augenbewegungen			
Anzahl	DOF	Ort	Aktionen
2	1	Auge	Augen rechts/links bewegen
1	1	Auge	Auge hoch/runter bewegen
	3	Menschenähnliche Augenbewegungen beim Verfolgen des Fokus of Attention(FoA), ermöglicht dem gegenüber Rückschluss auf Kismets Kismets Blickrichtung und damit des FoA	

Kopfbewegung			
Anzahl	DOF	Ort	Aktionen
1	3	Nacken	Kopf bewegen: rechts/links, nicken, vor/zurück
	3	Vereinfacht das Verfolgen des FoA, ermöglicht eine begrenzte Bewegung auf den FoA zu/weg (signalisiert Interesse/Furcht)	

2.1.2 Sensorik: visuell

Kismet sieht mittels 4 Kameras. Auf der „Nase“ zwischen den Augen sowie direkt über der Oberlippe sitzen zwei Weitwinkelkameras welche Kismet ein größeres Gesichtsfeld ermöglichen. Sie werden zur Ermittlung des FoA sowie zur Abstandsmessung (Stereobildmatching) verwendet. Die beiden Augenkameras verfolgen den FoA und übermitteln ein höher Aufgelöstes (vergrößertes) Bild des FoA. Auf diesen Bilddaten wird zum Beispiel die Augenerkennung durchgeführt um den FoA des Gegenüber zu ermitteln.

2.1.3 Sensorik: auditiv

Da Kismets Motoren sehr viele Störgeräusche produzieren sind Kismets „Ohren“ an den Interakteuren angebracht, diese Funkmikrophone übertragen ihre Aufnahmen direkt an die mit der Bearbeitung betrauten Rechner. Zur Zeit wird noch keine Worterkennung auf den Audiodaten durchgeführt sonder über Tonhöhe, Lautstärke und den zeitlichen Verlauf der Parameter auf die Intention des Sprechenden geschlossen (laut,hoch $\approx$ anschreien, normal+moduliert $\approx$ beruhigend usw.)

2.1.4 Akkustik

Kismet kann Lautäußerungen produzieren die so klingen, als ob ein kleines Kind sie produziert. Sie werden mittels einer Software so verändert das sie zum emotionalen Zustand Kismets passen (DECtalk v4.5, basierend auf Klatt Synthesizer).

2.2 Computer

Die im Netzwerk parralell arbeitenden Computer müssen folgende Anforderungen erfüllen:

- Echtzeitverarbeitung der Video-Daten (30 Bilder pro Sekunde), Latenzzeit max. 500ms
- Echtzeitverarbeitung der Audiosignale (8kHz), Latenzzeit max. 500ms

Sind die Sensordaten eingegangen sind folgende Aufgaben zu bewältigen:

- Bilddaten analysieren

- Tondaten analysieren
- FoA bestimmen
- Motivations- und Verhaltenssystem abgleichen
- Expressive Aktionen: Motorsteuerug des Gesichts/Kopfes, Sprachausgabe

Dabei darf die Reaktion auf einen Stimulus nicht zu lange dauern, damit für den Menschen ein kausaler Zusammenhang zwischen seiner Handlung und Kismets Reaktion gewahrt bleibt.

Geleistet wird dies alles durch 15 Computer:

9 400MHz PCs (QNX): Motor Kontroller, Attention System, Target Tracker, Target Distance, Skin filter, Motion Filter, Color filter, Eye finder, Audio data

4 Motorola 68332 (L): Face Control, Perception & Motor, Emotive Response, Drives&Behaviour

1 Dual 450 MHz PC (NT): speech synthesis, vocal affective intent recognition

1 500 MHz PC (Linux): speech recognition

QNX = Echtzeit Unix, L = selbstentwickeltes multithreaded Lisp

3 Software

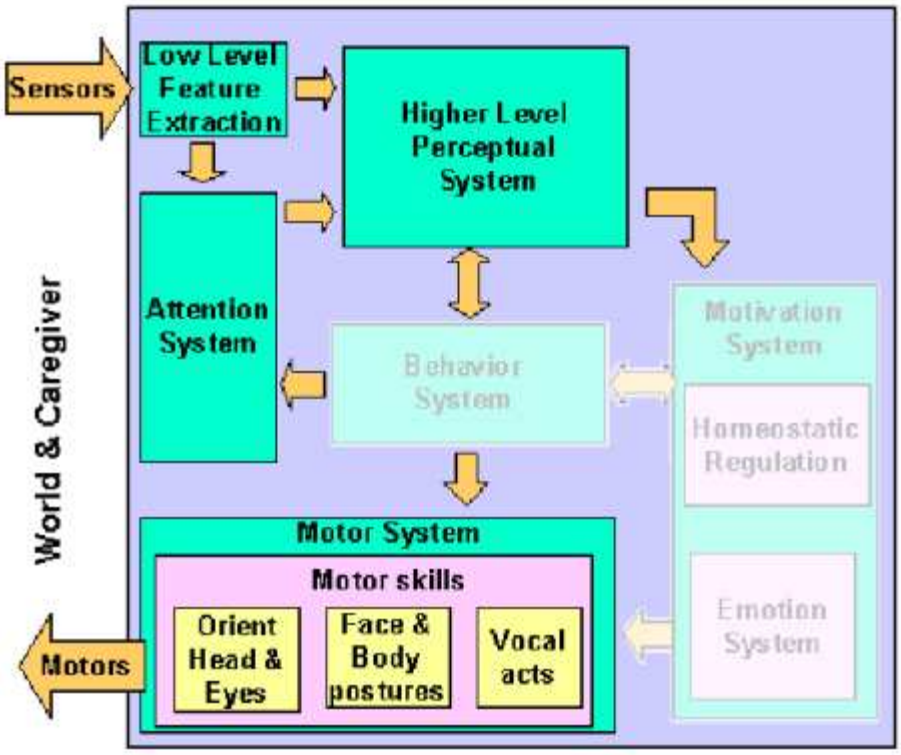

Abbildung 4: Aufteilung der Software

Kurzbeschreibung: In der *Low Level Feature Extraction* werden die eingehenden Bilder vorverarbeitet und verschiedene saliency maps berechnet. Auf diesen werden vom *Attention System* verschiedene FoA berechnet. Diese werden zusammen mit den Bildern ans *Higher Level Perceptual System* weitergeleitet welches dann diese Inputs in Kombination mit dem derzeitigen Zustand Kismets auswertet und einen Verhaltensprototypen an das *Behaviour System* und das *Motivation System* weitergibt. Diese stoßen dann die Antwort Kismets an, indem dem *Motor System* die Expressiven Parameter für die Motoren mitgeteilt werden. Die Überblendung von vorherigen in den jetzigen Motorzustand wird dort berechnet und angesteuert.

3.1 Vorverarbeitung der Bilddaten

Aus den Bilddaten der Weitwinkelkameras werden verschiedene Saliency-Maps berechnet. Aus der Kombination dieser Maps wird der FoA festgelegt. Anhand der Bilddaten der

Augenkameras vom FoA werden weitergehende Features extrahiert (zB Gesichter, ...). Die Saliency Maps sind an menschlichen Kleinkindern orientiert: helle/leuchtende Farben, hautfarbige Gebiete, Bewegung, Distanz zum Ziel, Distanz zum alten FoA, usw.

- helle/leuchtende Farben: rot, blau, grün, gelb
- Hautfarbe
- Bewegung
- Augen
- Distanz zum Ziel
- große Gegenstände / starke Bewegung (≈Bedrohung)
- (Distanz des neuen möglichen FoA zum letzten FoA)

3.1.1 Farben

Die RGB-8bit-Videodaten werden in ein vier-Farb-Modell (RGBY)umgerechnet und dabei anhand der gewichteten Helligkeit der RGB-Werte normalisiert. Heraus kommt eine saliency map die helle/leuchtende Farbwerte hervorhebt.

3.1.2 Hautfarbe

Aus dem Eingabebild werden anhand einer bestehenden Hautfarbtabelle mögliche Haut-Gebiete herausgefunden. Vorhergehende Hautfarben werden in die Berechnung der jetzigen Hautfarbe mit einbezogen.

3.1.3 Bewegung

Aus dem Bild des letzten Zeitschritts der Weitwinkelkameras und dem jetzigen Bild wird eine Verschiebung berechnet. Diese Verschiebung wird aus den Weitwinkelbildern berechnet, da der Kopf meist still steht und nur die Augen den FoA verfolgen, die so erhaltene saliency map wird weichgezeichnet. Helle Stellen auf dieser Karte korrespondieren mit viel Bewegung, dunkle stellen ohne Bewegung.

Um Bewegungsartefakte durch Eingenbewegung auszuschließen ist die Bewegungsdetektion während Eigenbewegungen deaktiviert.

3.1.4 Augen

Mittels Vorbedingungen werden Augen gesucht:

- Augen sind dunkle Regionen in einem Gebiet von Hautfarbe
- zwischen Augen liegt der helle Nasenrücken
- über den Augen liegt die helle Stirn
- der Kopf ist nur wenig verdreht, die Augen liegen horizontal in einer Höhe
- Mensch ist in Interaktionsdistanz (0.9-2m)

(Developed by Aaron Edsinger (edsinger@ai.mit.edu))

3.1.5 Distanz zum Ziel

Der FoA auf den Stereobildern der Weitwinkelkameras werden durch Verschiebung in Übereinstimmung gebracht. Dies geschieht für mehrere Punkte im FoA. Aus der bekannten Robotergeometrie wird die Entfernung zum Ziel berechnet

3.1.6 große Gegenstände / starke Bewegung (≈Bedrohung)

Können die beidem Weitwinkelbilder nicht in Übereinstimmung gebracht werden, so spricht dies dafür das ein sehr naher Gegenstand die Kamera blockiert, das heisst jemand in den Privatbereich Kismets eindringt.
Berechnet wird dies durch eine einfache Pixeldifferenz der beiden Stereobilder. Eine große Differenz ist gleichbedeutend mit einem großen, nahen Objekt, also einer möglichen Bedrohung.

Ergibt sich aus der Bewegungsdetektion eine hohe Bewegungsrate, „erschrickt“ er ebenfalls.

Beides führt zu einem Rückzug Kismets und damit einem Signal das dieses zu schnell/dicht war.

3.2 Verarbeitung der Audiodaten

Die Sprachverarbeitung (MIT, Spoken Language Systems Group (www.sls.lcs.mit.edu/sls)) von Kismet registriert in Echtzeit den zeitlichen Verlauf der Tonhöhe und Lautstärke der Audiosignale und unterscheidet Sprache von Umgebungsgeräuschen:

- sound present
- speech present
- time stamped Tonhöhe
- time stamped Lautstärke
- time stamped Phoneme

Im Moment beeinflussen Lautstärke und Tonhöhe Kismets emotionalen Zustand.

3.3 Aufmerksamkeitssteuerung

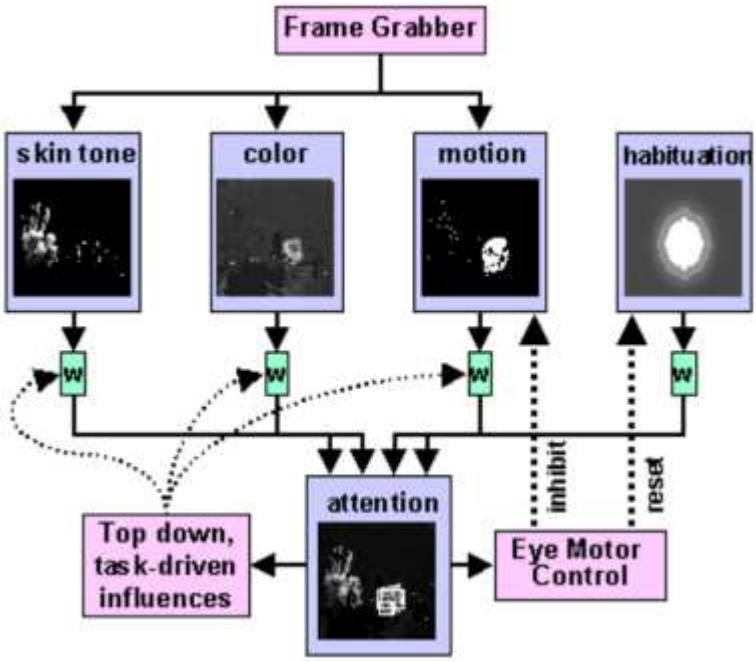

Abbildung 5: Saliency maps

Kismets reagiert auf ähnliche Stimuli wie ein Mensch, dies ist notwendig damit der Benutzer leicht Kismets Aufmerksamkeit erregen und erhalten kann. Die Echzzeitberechnung des FoA ist wichtig um Kausalität zwischen Aktion/Reaktion zu erhalten. Kismet signalisiert sein Interesse durch Mimik, Folgebewegungen der Augen und des Kopfes. Dabei nutzt er menschenähnliche Augenbewegungen um Dinge zu verfolgen und seine Umgebung zu erforschen.
Die Aufmerksamkeitssteuerung basiert auf dem *Guided Search System* von Wolfe (ein Modell für menschliches visuelles Sehen) erweitert um Kamerasteuerung und Rückkopplung mit dem Motivationssystem von Kismet.

Wie oben bereits ausgeführt Reagiert Kismet auf helle/leuchtende Farben, Bewegung, Größe der Objekte und Abstand zu Kismet. Dabei sind Stimuli nahe am derzeitigen FoA eher in der Lage Kismet FoA zu erlangen als Gegenstände mit größerem Abstand.

3.4 Motor System

Das Motor System erhält ein Ziel und eine Verhaltensstrategie, entscheidet welche Elementarbewegungen nötig sind (Körperhaltung, Blickrichtung, Sprache, Gesichtsausdruck) und verschmilzt diese Einzelsequenzen miteinander.

4 Zukunft

- Optimierung der visuellen, auditorischen und motorischen Steuerung in Echtzeit
- Erweiterung von Kismet auf ein selbstlernendes System welches durch Interaktion mit verschiedenen Benutzern von diesen lernt
- Das Beibringen soll für den Benutzer einfach und intuitiv erfolgen, Hinweise und Verhaltensweisen aus der Mensch-Mensch-Kommunikation soll Kismet „verstehen“ und darauf adequat reagieren (und sie als Feedback auch reproduzieren) um die Interaktion zu ermöglichen
- Dies erfordert Nachdenken über Themen wie: Identität, Gedächnis, Selbst- und Fremderkenntnis, Absichten, Gefühle, Soziales lernen (Imitation), Ethik

5 Quellen

- Homepage zu Kismet
  (http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html)