

Studienarbeit

Kontext im intelligenten Raum:

Visuelle Klassifikation von Aktivitäten

Universität Karlsruhe (TH)
Fakultät für Informatik
Institut für Logik, Komplexität und Deduktionssysteme
Prof. A. Waibel

Autor: Klaus Fritzsche
Betreuer: Dr. R. Stiefelhagen

Datum: 05.10.2004

Inhaltsverzeichnis

1	Einleitung	2
1.1	Der intelligente Raum	3
1.2	Was ist Kontext?	3
1.3	Mögliche Zielanwendungen	4
2	Bestehende Ansätze	6
3	Anwendungsszenario	9
3.1	Datenbasis	9
3.2	Synchronisation	10
3.3	Klassifikationsziele	11
4	Klassifizierung von Aktivitäten mit KNN	13
4.1	Verarbeitung der Bilder	13
4.1.1	Die Bildfunktion	14
4.1.2	Betrachtung von Teilbildern	14
4.2	Der Merkmalsraum	14
4.2.1	Wahl der Merkmale	15
4.2.2	Bewegung	16
4.2.3	Historie von Differenzbildern	17
4.2.4	Hautfarbe	18
4.2.5	Optischer Fluss	20
4.2.6	Intensität	21
4.3	Klassifikation	21
4.3.1	Zusammensetzung der Eingabevektoren	22
4.3.2	Auswahl und Aufbau eines Klassifikators	22
4.3.3	Weiterverarbeitung der Ausgaben	24
4.3.4	Alternative Klassifikatoren	24
5	Experimente	25
5.1	Datenbasis	25
5.2	Training der Klassifikatoren	25
5.3	Ergebnisse	26

<i>INHALTSVERZEICHNIS</i>	1
5.3.1 Ergebnisse ohne Betrachtung von Teilbildern	26
5.3.2 Ergebnisse mit 4x4 Teilbildern	27
5.3.3 Ergebnisse mit 5x5 Teilbildern	28
5.3.4 Diskussion	28
6 Zusammenfassung und Ausblick	30
Tabellenverzeichnis	31
Abbildungsverzeichnis	31
Literaturverzeichnis	32

Kapitel 1

Einleitung

Der Bedarf und das Interesse an anwendungsfreundlichen, aufmerksamen und umsichtigen Informationssystemen ist enorm und beschäftigt verschiedenste Domänen in der Forschung. Trotz der Fortschritte, die auf dem Gebiet der künstlichen Intelligenz in den letzten Jahrzehnten erzielt wurden, ist der Gebrauch von Computern immer noch eine mühselige und unnatürliche Angelegenheit. Das stille Sitzen vor dem Bildschirm sowie die Eingabe bestimmter Abfolgen von Tastaturkommandos prägt noch immer die Benutzung eines solchen. Die Interaktion mit dem Computer beansprucht so ein hohes Maß an Aufmerksamkeit des Benutzers und erfordert gleichzeitig ein gutes Verständnis des Systems mit seiner Benutzeroberfläche sowie den Ein- und Ausgabemöglichkeiten.

Auch wenn dies in vielen Anwendungen nur schwerlich zu umgehen ist - man denke beispielsweise an die Arbeit mit Office-Programmen oder das Programmieren von Anwendungen - so gibt es doch viele Anwendungsbereiche, in denen eine innovativere, intelligentere Art und Weise wünschenswert wäre, mit einem System zu interagieren oder dieses mit Informationen zu versorgen. Diese Ansätze fasst der Begriff *pervasive¹ computing* zusammen. Angestrebt werden Dienste, die zu jeder Zeit und an jedem Ort verfügbar sind und sich an die aktuelle Situation anpassen können. Im Idealfall unterstützen diese Systeme den Menschen bei all seinen täglichen Aktivitäten in einer Weise, die ihn vergessen lässt, es überhaupt mit Computern zu tun zu haben.

Die Erschließung dieser Anwendungsbereiche ist jedoch noch nicht weit fortgeschritten. Die Forschung ist noch in den Anfängen und es gibt nur wenig praxistaugliche Lösungen. Ubiquitäre Systeme, die Situationen begreifen können und selbstständig unterstützend auf diese einwirken können, liegen immer noch fern in der Zukunft. Im Vordergrund steht in diesen Systemen nicht mehr die Mensch-Maschine Interaktion als vielmehr die Beobachtung der Aktivitäten und der Interaktionen *zwischen* den Menschen.

Eine Bündelung der Forschungsarbeiten in diesem Bereich stellt das EU-Projekt CHIL

¹deutsch: durchdringend, überall vorhanden, tief greifend

dar (Computers in the Human Interaction Loop, <http://chil.server.de>), in dem verschiedene Einrichtungen Europas gemeinsam an dieser Vision arbeiten. Die Forschungsgruppen arbeiten an modernen Systemen, die ihre Dienste unauffällige im Hintergrund anbieten sollen und möglichst wenig direkte Interaktion und somit Störung des Benutzers verursachen.

1.1 Der intelligente Raum

Der so genannte intelligente Raum (englisch: Smartroom) spielt eine wichtige Rolle in diesem Forschungsbereich. Es handelt sich hierbei um einen mit umfangreicher Sensorik ausgestatteten Raum, in dem Meetings, Seminare oder Präsentationen abgehalten werden können. In diesem wohlbekanntem, abgesteckten Bereich werden Verfahren entwickelt und Grundlagen gelegt, die auch in anderen Szenarien Anwendung finden.

Ziel ist es, den Computer immer mehr in den Hintergrund der menschlichen Aufmerksamkeit rücken zu lassen. Die Geschehnisse im Raum sollen erfasst und darauf reagiert werden. Eine Vielzahl der Dienste soll selbstständig auf Basis der aus dem Raum gewonnenen Informationen abgeleistet werden. Die Interaktion mit dem Menschen soll hierbei auf ein Minimum reduziert werden. Vielmehr soll die Beobachtung der menschlichen Aktivitäten und der Mensch-zu-Mensch Kommunikation im Mittelpunkt stehen.

Auch in dieser Arbeit wurde der intelligente Raum als Zielszenario gewählt.

1.2 Was ist Kontext?

Kern vieler Anwendungen im intelligenten Raum ist das sensorische Überführen von Umgebungsdaten in eine interne Repräsentation und das Gewinnen einer semantischen oder besser: pragmatischen Bedeutung aus dieser Information.

Der Kontext in einem Szenario zeichnet sich typischerweise durch seine Unabhängigkeit von speziellen Anwendungsdomänen aus. Antworten auf die Fragen Wer?, Was?, Wann?, Wo?, Wie? oder Warum? formen eine Wissensbasis, die für viele Anwendungen von entscheidender Bedeutung ist. Oftmals wird dieser Kontext ignoriert, da für dessen Auswertung Merkmale modelliert und ausgewertet werden müssen, die nicht direkt mit der eigentlichen Anwendungsdomäne zusammenhängen.

Alex Pentland verdeutlicht den Sachverhalt in [CP00] mit der Betrachtung eines typischen Klassifikators. Diesen reduziert er auf die Berechnung einer bedingten Wahrscheinlichkeitstabelle $\mathcal{P}(y | u)$. Die zu treffende Entscheidung des Klassifikators wird hier durch y repräsentiert. u ist der Vektor der verwendeten Merkmale, die dem Klassifikator für seine Entscheidung zugrunde liegen.

Typischerweise gibt es eine große Zahl von Merkmalen v , die einfach ignoriert werden. Den Optimalfall würde natürlich die Berechnung von $\mathcal{P}(y|u, v)$ widerspiegeln, welche alle theoretisch zur Verfügung stehenden Merkmale in Betracht zieht. Da aber v unübersehbar groß ist, fällt die Wahl auf u - das Ergebnis einer aufgabenorientierten Merkmalsselektion.

Das Benutzen von Kontextwissen läßt sich ausdrücken als eine Komprimierung von v auf den Kontext c . Dies geschieht durch die Berechnung von $\mathcal{P}(c|v)$ mit einem handhabbar kleinen $|c|$. Dann läßt sich das Klassifikationsproblem abschätzen durch:

$$\mathcal{P}(y|u, v) \approx \mathcal{P}(y|u, c)$$

Ein großer Vorteil der Verfügbarkeit von Kontextwissen ist dessen Unabhängigkeit von der zu lösenden Klassifikationsaufgabe. So kann das Kontextwissen bei einer Vielzahl von Problemstellungen verwendet werden - zum Beispiel beim Berechnen einer anderen Wahrscheinlichkeitstabelle $\mathcal{P}(z|u, c)$. Verlässliches Kontextwissen ist somit durch seine Unabhängigkeit von der Anwendungsdomäne sehr vielseitig einsetzbar.

1.3 Mögliche Zielanwendungen

Das Erkennen von Aktivitäten und das Deuten von Kontextinformationen ist eine zentrale Aufgabe, die auf dem Weg zur natürlichen Interaktion mit einem Informatiksystem zu bewältigen ist. Auch und vielleicht besonders Anwendungen, in denen das Erkennen von Aktivitäten nicht die Kernanwendung ist, können und sollten den Kontext als wichtige Entscheidungsgrundlage anderweitiger Klassifikationen nutzen.

Beispielsweise könnte solch eine kontextbewusste Erweiterung einer bestehenden Anwendung den Mensch-Maschine-Dialog verbessern, indem sie Beobachtungen aus der Kommunikation zwischen den Menschen als erweiterte Entscheidungsgrundlage heranzieht und so Fehlinterpretationen vorbeugt. Das Begreifen von Situationen hätte eine robustere Deutung von Befehlen zur Folge, die nicht aus präzisen Eingabegeräten wie Maus oder Tastatur stammen. Die Ambiguität von Zeigegesten etwa oder die schwierige Extraktion von sprachlichen Befehlen aus einer Unterhaltung seien hier nur beispielhaft als Problembereiche angeführt, in denen ein solides Kontextwissen unersetzlich ist. Ein Problem bei jeglicher Analyse von Kommunikation zwischen Menschen liegt darin, dass die natürliche Kommunikation zwischen Menschen nicht immer eindeutig ist und somit nicht verlustfrei aus dem Kontext heraus genommen werden kann. Ähnliche Gesten beispielsweise können sich auf semantischer Ebene stark unterscheiden, wenn sie in verschiedenen Situationen beobachtet werden.

Andere Bereiche, in denen die Klassifikation von Aktivitäten schon jetzt eine Rolle spielt, sind zum Beispiel die automatische Verkehrsüberwachung, die Überwachung großer Menschenmengen oder der Schutz von kritischen Bereichen wie Geldautomaten.

Auch die Unterstützung von Akteuren im intelligenten Raum ist insbesondere im Rahmen dieser Arbeit wichtig zu erwähnen. So ist es denkbar, bestimmte Ereignisse wie Seminare durch Starten und Stoppen der Kameras und Mikrofone automatisch aufzunehmen. Weitere Beispiele wären die automatische Regelung der Lichtverhältnisse oder der automatische Start eines Übersetzungssystems in mehrsprachigen Unterhaltungen oder Seminaren.

Eine weitere Möglichkeit besteht darin, die gewonnenen Kontextinformationen einfach aufzuzeichnen. Eine solche Aktivitätenanalyse liefert eine semantische Repräsentation der Geschehnisse im beobachteten Szenario. Sie bietet beispielsweise die Möglichkeit, sich einen schnellen Überblick über eine große Menge von rohen Videodaten zu verschaffen, ohne die Daten im Einzelnen betrachten zu müssen.

Kapitel 2

Bestehende Ansätze

Das Erkennen und Deuten menschlicher Aktivitäten ist in verschiedenen Bereichen und unterschiedlichsten Umgebungen von Interesse. So beschäftigen sich derzeitige Arbeiten mit Systemen zur automatischen Verkehrsbeobachtung, mit intelligenter Überwachung oder intuitiver Robotersteuerung. Auch Büroumgebungen sind immer wieder Zentrum des Interesses.

Brian Clarkson und Alex Pentland modellieren in [CP00] den Kontext von Benutzern tragbarer Computer in Büroumgebungen, beim Einkauf in der Stadt und im Supermarkt. Die verwendete Sensorik ist hier darauf ausgelegt, die periphere Wahrnehmung des Menschen von seiner Umgebung zu simulieren. Die so entstehenden Merkmale werden als "non-attentional" bezeichnet, also als nicht zielgerichtet oder nicht auf etwas bestimmtes achtend. Sie sind nicht darauf ausgelegt, ganz bestimmte Situationen zu erfassen oder beispielsweise bestimmte Objekte zu erkennen. Im Gegensatz zu der Sichtweise, die eine Auswertung der Daten eines Nahsprechmikrophons oder einer in Blickrichtung der Person gerichteten Kamera liefern würde, kommt es hier eher auf den allgemeinen Eindruck von der Umgebung und die natürliche Situation des Handelnden an. Die zeitliche Granularität der betrachteten Klassen geht über die unmittelbare Situation hinaus und erfasst auch den Bereich längerer zeitlicher Abfolgen.

In den dokumentierten Experimenten kommen sowohl überwachte als auch unüberwachte Lernverfahren zum Einsatz. In den Experimenten mit überwachten Lernverfahren wurden audio-visuelle Daten in einen Merkmalsraum mit 24 Dimensionen abgebildet. Für die Unterscheidung von sechs verschiedenen Aktivitätsklassen, die sich hauptsächlich auf die Aufenthaltsorte der Testpersonen bezogen, wurde für jede der Klassen ein binärer Erkenner in Form von eines HMMs trainiert.

Auch die Experimente mit unüberwachten Lernverfahren basierten auf audio-visuellen Daten. Hier kam unter anderem eine Weitwinkelkamera zum Einsatz, deren Bildsequenzen ähnlich wie in dieser Arbeit (siehe Kapitel 4) in Teilbildbereiche aufgeteilt wurden. Es wurde jeweils die Korrelation der mit K-Means gefundenen Cluster im Merkmalsraum mit

den per Hand gelabelten Daten untersucht.

Mit der Analyse von Aktivitäten in Meetings beschäftigt sich [GCBBB03]. Anhand von audio-visuellen Merkmalen und HMMs werden hier Klassen wie zum Beispiel *Es finden ein Monolog statt*, *Notizen werden gemacht* und *Diskussion* voneinander unterschieden. Als visuelle Merkmale dienen Hintergrund- und Hautfarbinformationen. Da es nach Meinung der Autoren bei der Erkennung solcher Aktivitätsklassen oft nicht genau auf die zeitliche Abgrenzung derer, sondern viel mehr auf die Rekonstruktion der richtigen Reihenfolge ankommt, schlagen diese vor, zur Leistungsbestimmung derartiger Systeme eine an die Spracherkennung angelehnte Metrik vor, die so genannte *action error rate*. Diese berechnet sich aus der Anzahl der vom Erkennen ausgelassenen, hinzugefügten sowie ersetzten Aktivitäten in einer Sequenz. In den durchgeführten Experimenten wurde eine *action error rate* von insgesamt 20% erreicht.

Eine visuelle Unterscheidung der Klassen *keine Person*, *eine Person*, *eine aktive Person* und *mehrere Personen im Büro* wird in [OHG] mit einem auf geschichteten HMMs (LHMMs) basierenden Ansatz angestrebt. Als Merkmale dienen Hautfarb- und Bewegungsevaluatoren sowie eine Hintergrundsubtraktion und ein Gesichtsdetektor. Eine noch größere Anzahl von Klassen soll unter Zuhilfenahme audio-visueller Merkmale und dem Aufzeichnen von Tastaturaktivitäten handhabbar werden. Ausprobiert wurde dies anhand der Aktivitäten *Telefonat*, *Präsentation*, *Konversation*, *Beschäftigte Person*, *Distanzkonversation* und *leeres Büro*. Hierbei stellte sich heraus, dass LHMMs in diesem Bereich robuster funktionieren und leichter zu trainieren sind, als einfache HMMs.

Zur Evaluation komplexer Aktivitäten wie zum Beispiel der Interaktion zwischen Personen kamen auch schon parametrische HMMs (PHMM) in [WB98], Entropic-HMMs in [BK00], HMMs variabler Länge (VHMM) in [GJH01] und verbundene HMMs (CHMM) in [BOP96] zum Einsatz.

Ein unüberwachtes Clusteringverfahren findet in [PJ] Anwendung. Mit einer dieser Arbeit ähnlichen Aufteilung der Bilder in einzelne Bereiche und der Auswertung von Farbton und Sättigung in den hier 16 verwendeten Kästchen werden verschiedene Aufenthaltsorte einer Person, die mit einem tragbaren Computer ausgestattet ist, unterschieden.

Neben Farbton und Sättigung kamen in verschiedenen Arbeiten auch noch andere Merkmale zum Einsatz, die in einzelnen Teilbildbereichen ausgewertet werden können: So verwenden Polana und Nelson in [PN94] die Summe der optischen Flussvektoren. Yamamoto *et al.* bestimmen in [YOI92] die Anzahl der Vordergrundpixel. Takahashi *et al.* definiert in [TSKO94] einen durchschnittlichen Kantenvektor für jedes Teilbild.

Neuronale Netze kommen in [LAL01] zur Erkundung des Kontextes zum Einsatz. Es werden die Vorteile einer Verwendung vieler kleiner Sensoren im Gegensatz zu einem komplexen Sensor herausgehoben. Kostenfaktoren, Robustheit und Flexibilität sprechen für die

Verteilung vieler kleiner Sensoren entweder im Raum oder (im Fall von tragbaren Computern) am Körper. Um die Vorverarbeitung der Sensordaten zu minimieren, werden einfache Merkmale wie Durchschnitt, Standardabweichung, Minimum und Maximum der Werte in einem bestimmten Zeitfenster berechnet. Ausgabe des Klassifikators ist eine Liste aller gelernten Kontextklassen mit deren Auftrittswahrscheinlichkeit im aktuellen Zeitfenster. So können auch Situationen gehandhabt werden, in denen sich verschiedene Klassen überlappen. Auch ein clusterbasierter Ansatz wird diskutiert, da man mit diesem flexibler ist, wenn es darum geht, im Nachhinein noch Kontext- oder Aktivitätsklassen hinzu zu fügen.

Kapitel 3

Anwendungsszenario

Als Ort unseres Anwendungsszenario stand ein Institutsseminarraum im ILKD¹ der Universität Karlsruhe (TH) zur Verfügung, in dem sich einige Arbeitsplätze befinden. Neben den Seminaren, die hier wöchentlich stattfanden, wurde der Raum an den restlichen Tagen von Studenten und Mitarbeitern des Instituts zur individuellen Arbeit genutzt.

3.1 Datenbasis

In den Ecken des Raumes sind Kameras fest installiert, deren Ausrichtung die gesamte Versuchsreihe über fixiert war. Jede der Kameras ist über Firewire mit einem PC verbunden.

Seminarveranstaltungen wurden - auch für andere Zwecke - mit einer Auflösung von 640x480 Pixeln und 15 Bildern pro Sekunde (FPS) aufgenommen. Insgesamt standen ca. 21 Stunden Seminarvideos zur Verfügung. Um ein möglichst realistisches Abbild des Alltags in dem Raum außerhalb der Seminarveranstaltungen zu erhalten, wurde mit den Kameras ohne gezielte Vorauswahl von Situationen ein größerer Zeitraum aufgenommen. Die Aufnahmen fanden sowohl nachts als auch in Hochbetriebszeiten statt. Aus Leistungs- und Platzgründen geschah dies mit einer Auflösung von 320x240 Pixeln und einer Bildwiederholrate von 3,75 FPS. Hierbei sind Aufzeichnungen in einer Länge von ungefähr 128 Stunden entstanden. Insgesamt entstand ein Datenvolumen von ungefähr 800 Gigabyte.

Eine Beispielaufnahme der vier Kameras ist in Abbildung 3.1 gezeigt.

¹Institut für Logik, Komplexität und Deduktionssysteme



Abbildung 3.1: Beispielaufnahme der vier Kameras

3.2 Synchronisation

Bei der Synchronisation der verschiedenen Kameras kam eine Middleware namens NIST² Smartflow [NIST] zum Einsatz, die das gemeinsame Starten und Stoppen von Aufnahmen ermöglichte. Auch eine zentrale Verwaltung der Kameraeinstellung war mit Hilfe der Middleware möglich.

Für die spätere Auswertung der Daten ist allerdings eine Synchronisation der Videostreams auch auf Einzelbildebene nötig. Zu diesem Zweck wurden die Videodaten in einem eigenen Videoformat gespeichert und jedes der Bilder mit einer genauen Zeitmarke der Aufnahme versehen. Bei der Auswertung wird dann eine der Sequenzen als Steuersequenz verwendet. Zu jeder Zeitmarke der Steuersequenz wurde jeweils der am besten passenden Frame in den drei restlichen Videosequenzen gesucht. So konnte eine ausreichende Zeitnähe zwischen den betrachteten Frames gewährleistet werden. Zeitfenster, in denen die Differenz der einzelnen Zeitmarken zu groß war, wurden verworfen.

In den folgenden Formeln wird folgende formelle Notation verwendet: Es liegen vier Bildsequenzen $b_1^i, b_2^i, b_3^i, \dots$ ($i = 1, \dots, 4$) mit den dazu gehörigen Zeitmarken $t_1^i, t_2^i, t_3^i, \dots$ ($i = 1, \dots, 4$) vor. Zur Synchronisation werden die Zeitmarken $t_1^s, t_2^s, t_3^s, \dots$ einer Sequenz s als Steuersequenz genutzt.

²National Institute of Standards and Technology, <http://www.nist.gov>

Die Funktion $find : N \times Z \rightarrow Z$ findet dann nach folgender Regel zu gegebener Zeitmarke t den am besten geeigneten Frame j in der Sequenz i :

$$find(i, t) = \begin{cases} \operatorname{argmin}_k (|t_k^i - t|) & \text{falls } |t_{\operatorname{argmin}_k (|t_k^i - t|)}^i - t| \leq \varepsilon \\ -1 & \text{sonst} \end{cases}$$

ε steht hier für die maximal erlaubte Differenz zweier Zeitmarken. Das zwei- bis dreifache des rechnerisch von der Framerate abhängigen Zeitabstandes zweier Frames hat sich hier als sinnvoller Wert erwiesen.

Es sind vier *synchrone* Bilder $b_{find(1,t)}^1, b_{find(2,t)}^2, b_{find(3,t)}^3, b_{find(4,t)}^4$ aus den verschiedenen Sequenzen zum Zeitpunkt t gefunden, wenn für alle $i = 1, \dots, 4$ gilt: $find(i, t) \neq -1$. Zeiträume, in dem sich keine synchronen Bilder finden, werden nicht betrachtet.

3.3 Klassifikationsziele

Unabhängig von den ohne Vorauswahl aufgenommenen Daten galt es festzulegen, welche Art von Aktivitäten überhaupt erkannt werden sollen. Dies richtet sich zum einen natürlich danach, was der Raum an Aktivitäten hergibt, zum anderen aber auch danach, was davon überhaupt erkannt und klassifiziert werden soll.

Ein Brainstorming hat eine Vielzahl von verschiedenen Aktivitäten ergeben, die in dem Raum auftreten könnten. Einige Beispiele sind:

- Ein Seminar findet statt
- Eine Person läuft durch den Raum
- Tür geht auf
- Jemand tippt auf einer Tastatur
- Zwischenfrage in einem Seminar
- Nichts passiert
- Drucker druckt

Allein dieser Listenauszug birgt eine Vielzahl von verschiedenen Klassen und damit potentiellen Klassifikationszielen. Um dieser Vielzahl Herr zu werden, wurden die gefundenen Aktivitäten in Gruppen unterschiedlicher zeitlicher und räumlicher Granularität aufgegliedert. Die entstandene Einteilung ist in Tabelle 3.1 beschrieben.

Bei Aktivitäten längerer zeitlicher Ausdehnung werden also lokale und globale Zustände

	auf einen Ort beschränkt	gesamten Raum betreffend
längere zeitliche Ausdehnung	"lokaler Zustand"	"globaler Zustand"
kurze zeitliche Ausdehnung	"lokales Ereignis"	"globales Ereignis"

Tabelle 3.1: Gruppeneinteilung möglicher Aktivitäten

unterschieden. "Lokal" bedeutet, dass die Aktivität auf einen bestimmten Ort im Raum beschränkt ist, an dem beispielsweise eine Person arbeitet. Globale Zustände beziehen sich auf den ganzen Raum und beschreiben somit Aktivitäten wie Seminarveranstaltungen oder allgemeine Laborarbeit.

Aktivitäten, die sich nur über kürzere Zeiträume erstrecken, wurden Ereignisse genannt. Auch hier werden die lokalen von den globalen Ereignissen unterschieden, die den ganzen Raum betreffen. Ein lokales Ereignis wäre zum Beispiel *Die Tür geht auf*. Global hingegen ist beispielsweise die Klasse *Das Licht geht an*.

Auch in [CP00] wird eine ähnliche Gruppierung in so genannte "scenes" und "events" vorgenommen. Wie diese Abgrenzung genau definiert ist, ist in anderem Kontext in [CP99] nachzulesen.

Die Experimente in dieser Arbeit konzentrieren sich hauptsächlich auf die so genannten Zustände. Eine Klassifikation dieser lang andauernden Klassen, die den gesamten Raum betreffen, schien besonders interessant, da Vorgänge dieser Art auf einer besonders hohen semantischen Ebene angesiedelt sind und die grundlegenden Zustände des Raumes widerspiegeln. Der gewählte Ansatz ist jedoch nicht auf diese Art von Aktivitäten beschränkt. Experimente mit den restlichen Gruppen von Aktivitäten sind zwar in dieser Arbeit nicht enthalten, zählen aber mit zu den unmittelbar nächsten Schritten in der Zukunft. Innerhalb der Gruppe der Zustände wurden folgende Klassen unterschieden:

- Seminar findet statt
- Laboralltag
- Sonstiges (Durchgangsverkehr, Gebäudereinigung, Raum leer, usw.)

Diese Einteilung spiegelt die realen Zustände des intelligenten Raums wieder. Seminarveranstaltungen und der Alltag im Labor sind typische Klassen für Aktivitäten, deren Erkennung bzw. Unterscheidung sehr hilfreich wäre. Eine weitere Unterteilung der Klassen in Unterklassen wie z.B. "Personen arbeiten am Tisch in der Mitte des Raumes" oder "Eine Person betritt den Raum" ist möglich. Diese Unterklassen fallen dann in die Gruppe der Vorgänge oder Ereignisse.

Kapitel 4

Klassifizierung von Aktivitäten mit KNN

Dieses Kapitel beschreibt die Verarbeitungsschritte von den rohen Videodaten bis hin zur Klassifikation. Anfänglich wird auf die Einteilung der Einzelbilder in Teilbildbereiche eingegangen. Daraufhin folgt eine Diskussion der Merkmale, welche in den Teilbildern extrahiert werden. Im letzten Abschnitt werden die verwendeten Klassifikatoren beschrieben.

4.1 Verarbeitung der Bilder

Die direkte Klassifizierung von Zuständen aus Videosignalen mehrerer Kameras stellt insgesamt ein großes Klassifizierungsproblem dar. Jedes der Pixel im Bild ist repräsentiert durch drei Farbwerte. Die Anzahl der Dimensionen des Merkmalsraumes wächst schon bei einer Betrachtung eines einzelnen Zeitfensters und einer Bildauflösung von 640x480 schnell in die Millionen. Will man einen ganzen zeitlichen Verlauf betrachten, so vervielfacht sich die Dimensionalität entsprechend schnell.

Für eine sinnvolle Abgrenzung semantischer Klassen in dem Material wird demnach eine aussagekräftige Vorverarbeitung der Daten benötigt. Diese muss die hohe Dimensionalität des Problems herunterbrechen, darf aber gleichzeitig nur mit möglichst wenig Verlust von semantischer Information verbunden sein. Als weitere Anforderung an eine geeignete Abstrahierung der Daten kommt hinzu, dass in einem späteren Zielsystem die Echtzeitfähigkeit der Vorverarbeitung entscheidend sein könnte. Die Vorverarbeitung darf also nicht zu rechenintensiv sein.

In dieser Arbeit werden einzelne Regionen der Bilder zusammenfassend ausgewertet - und zwar mit möglichst einfach zu berechnenden, elementaren Merkmalen, auf deren Wahl später noch näher eingegangen wird. Die Bilder werden in gleichmäßige, rechteckige Regionen aufgeteilt. Eine Zahl von 25 Feldern pro Kamerabild hat sich hier als sinnvoll erwiesen. Die Auswertungsergebnisse in den einzelnen Feldern dienen dann als Eingabevektoren für

den Erkennen.

4.1.1 Die Bildfunktion

Ein Bild b der Größe $r_x \cdot r_y$ wird repräsentiert durch seine zugehörige Bildfunktion b mit

$$b : (x, y) \mapsto (r, g, b),$$

wobei r, g und b jeweils die Farbwerte des Bildes im RGB-Farbraum $([0 \dots 255]^3)$ darstellen. Die Definitionsbereiche für x und y sind entsprechend der Auflösung $[0 \dots r_x - 1]$ und $[0 \dots r_y - 1]$.

4.1.2 Betrachtung von Teilbildern

Zur effizienten Verarbeitung der Bilder wurden die Einzelbilder in rechteckige Teilbilder aufgeteilt. In jedem der entstehenden Kästchen findet dann eine separate Auswertung statt.

Bei einer Aufteilung des Bildes b in k_x Kästchen in x -Richtung sowie k_y Kästchen in y -Richtung (siehe Abbildung 4.1) wird das i -te Kästchen von links und j -te Kästchen von oben repräsentiert durch die Funktion

$$B^{i,j}(b) \text{ mit } (x, y) \mapsto b(i \cdot r_x / k_x + x, j \cdot r_y / k_y + y),$$

welche eine Zuordnung der Pixel im Bildbereich zu den Pixeln im Ausgangsbild vornimmt. x und y sind hier jeweils definiert in den Bereichen $[0 \dots r_x / k_x - 1]$ und $[0 \dots r_y / k_y - 1]$, welche die Ausdehnung des Teilbildes beschreiben.

4.2 Der Merkmalsraum

Zur Auswertung der Bildinformationen wurden Merkmale verwendet, die bestimmte Eigenschaften eines Bildes (bzw. eines Teilbildes) zu einem Zeitpunkt oder innerhalb eines zeitlichen Verlaufs durch einen Zahlenwert zwischen 0 und 1 ausdrücken.

Sei \mathcal{B} die Menge aller zur Verfügung stehenden Bilder. Ein Merkmal f mit

$$f : \mathcal{B}^h \rightarrow [0 \dots 1] \subset \mathbb{R}$$

ist dann eine h -stellige Funktion, deren Stelligkeit h sich nach dem betrachteten Zeitraum

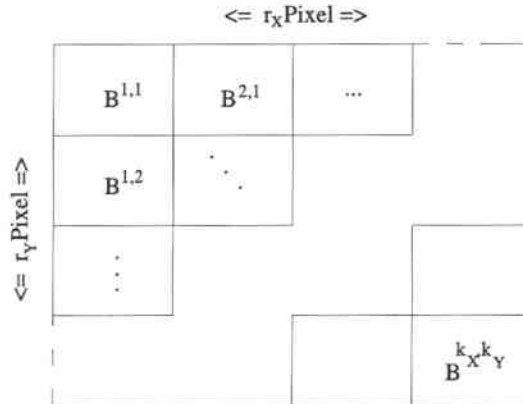


Abbildung 4.1: Aufbau eines Einzelbildes

richtet. Wird nur ein einzelner Zeitpunkt ausgewertet, so gilt $h = 1$. Wird ein zeitlicher Verlauf betrachtet, so gilt entsprechend $h > 1$ und es wird eine Historie von Bildern (bzw. Teilbildern) ausgewertet.

Die Merkmale betrachten jede der Bildsequenzen für sich. Mechanismen wie räumliche Triangulation von Punkten, die von mehreren Kameras gleichzeitig aufgenommen wurden, sind mit diesen Merkmalen nicht zu realisieren. Es bleibt dem späteren Klassifikator überlassen, derartige Zusammenhänge zwischen den Bildern zu finden. Eine gewisse räumliche Information bleibt auch in den Merkmalen kodiert, da ein fester Punkt im Raum immer in den gleichen Teilbildbereichen seine Auswirkung zeigt.

4.2.1 Wahl der Merkmale

Das Finden geeigneter Merkmale ist ein besonders wichtiger Aspekt der Implementierung dieses Ansatzes. Die Merkmale verringern die Dimension des Klassifikationsproblems enorm. Sie sollten eine gewisse Abstrahierung von den rohen Pixeldaten bieten und gleichzeitig die entscheidenden Informationen wahren. Außerdem findet bei der Auswertung ganzer Bildhistorien mittels mehrstelliger Merkmale auch die zeitliche Dimension ihre Beachtung. Die Betrachtung reiner ist-Zustände des Raumes wäre der Anforderung an den Ansatz nicht gewachsen, auch längerfristige Aktivitäten oder Zustände zu deuten. Der zeitliche Verlauf ist hier von zentraler Bedeutung.

Ein weiteres wichtiges Kriterium bei der Wahl geeigneter Merkmale ist deren Berechnungsaufwand. Ein zukünftiges System sollte die Klassifikation in Echtzeit bewältigen können. Zu komplizierte Berechnungen in den einzelnen Teilbildern wären hinderlich.

Im Folgenden werden nun einige Merkmale diskutiert, die dann auch experimentell ge-

testet wurden. Die Ergebnisse sind in Kapitel 5 dokumentiert.

4.2.2 Bewegung

Ein wichtiges Merkmal bei der Klassifikation von menschlichen Aktivitäten und Zuständen des Raumes ist die Bewegung im Bild. Die Bildung der Differenz von aufeinander folgenden Bildern gibt Aufschluss darüber, wo sich etwas verändert, also bewegt, hat. Eine Visualisierung dieses Merkmals zeigt Abbildung 4.2.

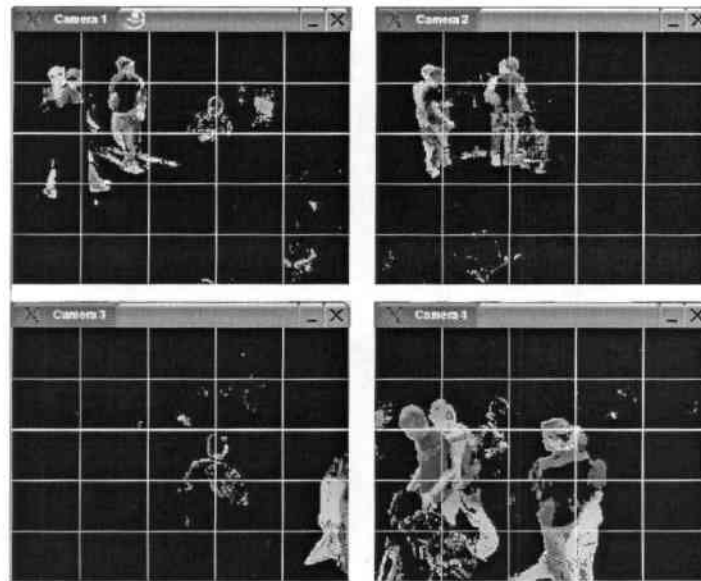


Abbildung 4.2: Differenzbild

Die Berechnung des Differenzbildes findet immer auf zwei aufeinander folgenden Teilbildern B_{i-1} und B_i einer Kamera statt. Zuerst wird Pixelweise die Differenz auf jedem der Kanäle berechnet. Ist die Norm des entstehenden Differenzvektors größer als ein vorher festgelegter Schwellwert ε , so wird das Pixel mit 1 (*Hier hat sich etwas bewegt*) bewertet. Für alle unter dem Schwellwert liegenden Pixel gibt die Funktion 0 (*Hier hat sich nichts bewegt*) zurück:

$$\text{diff}_{B_{i-1}, B_i}(x, y) = \begin{cases} 1 & \text{falls } \|B_{i-1}(x, y) - B_i(x, y)\| > \varepsilon \\ 0 & \text{sonst} \end{cases}$$

In den Experimenten hat sich ein Schwellwert $\varepsilon = 40$ als sinnvoll erwiesen. Die auf diese Weise mit 1 klassifizierten Pixel sind in Abbildung 4.2 erhalten geblieben, die restlichen wurden schwarz eingefärbt.

Das eigentliche Merkmal $f_{\text{Bwg}} : \mathcal{B}^2 \rightarrow [0 \dots 1] \subset \mathbb{R}$ berechnet sich dann im Falle der Teilbilder B_{i-1} und B_i folgendermaßen:

$$f_{\text{Bwg}}(B_{i-1}, B_i) = \frac{k_x \cdot k_y}{r_x \cdot r_y} \sum_{(x,y) \in B} \text{diff}_{B_{i-1}, B_i}(x, y)$$

Der berechnete Wert spiegelt den Anteil der Pixel im Teilbild wieder, die sich verändert haben. Die Konstanten $r_{x/y}$ und $k_{x/y}$ stehen für die Auflösung des Gesamtbildes und die Anzahl der Teilbilder in x- und y-Richtung.

4.2.3 Historie von Differenzbildern

Um die zeitliche Ausdehnung des Merkmals f_{Bwg} zu erhöhen und somit ganze Bewegungsabläufe in Betracht zu ziehen, besteht die Möglichkeit, in dem Teilbild eine Historie von Differenzbildern zu akkumulieren. Die Bewertung der Differenz auf Pixelebene im Sinne von *hier hat sich etwas bewegt* oder *hier hat sich nichts bewegt* geht dann über in eine kontinuierliche Funktion, die Auskunft darüber gibt, wie viel Bewegung das Pixel in einen definierten Betrachtungszeitraum erfahren hat. Ältere Bewegungsbeobachtungen fließen mit geringerer Gewichtung als aktuelle Beobachtungen in die Berechnung mit ein. Eine Visualisierung des Merkmals liefert Abbildung 4.3. Hier sieht man, wie eine durch den Raum laufende Person einen immer schwächer werdenden Schweif von Bewegungsbeobachtungen hinter sich her zieht. Vorteil dieses Merkmals ist auch, dass ein kurzzeitiges Stillhalten einer Person diese nicht gleich unsichtbar für das Merkmal macht.

Die Berechnung der Bewegungsabläufe kann rekursiv ablaufen. Die Bewertung eines Bildes B_i setzt sich zusammen aus der Bewertung des Teilbildes B_{i-1} und des aktuellen Differenzbildes:

$$\hat{f}_{\text{BwgHist}}(B_i) = f_{\text{BwgHist}}(B_{i-1}) \cdot \lambda + f_{\text{Bwg}}(B_{i-1}, B_i)$$

Um im Wertebereich $[0 \dots 1] \subset \mathbb{R}$ zu bleiben, ist das eigentliche Merkmal f_{BwgHist} definiert als:

$$f_{\text{BwgHist}}(B_i) = \begin{cases} \hat{f}_{\text{BwgHist}}(B_i) & \text{falls } \hat{f}_{\text{BwgHist}}(B_i) < 1 \\ 1 & \text{sonst} \end{cases}$$

$\lambda \in [0 \dots 1] \subset \mathbb{R}$ ist der Faktor, mit dem die alten Bewegungsdaten abgeschwächt werden. In den Experimenten wurde ein Faktor von $\lambda = 0,95$ verwendet.

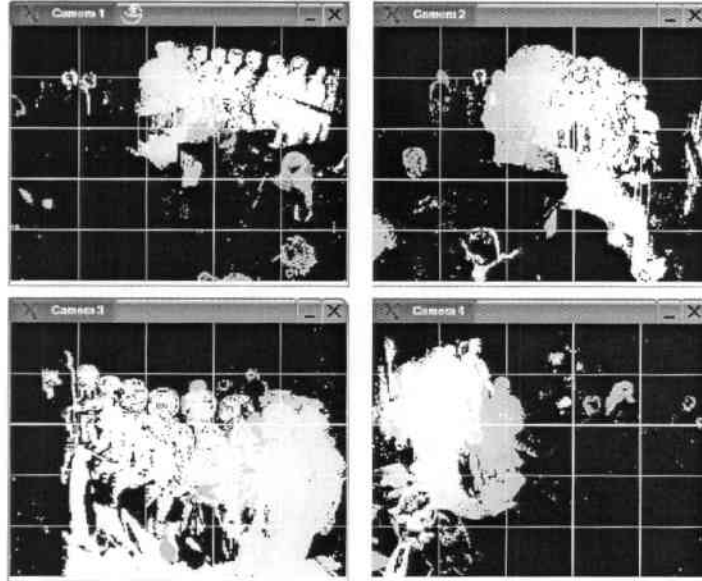


Abbildung 4.3: Bewegungsabläufe über die Zeit

4.2.4 Hautfarbe

Da die Erkennung von Aktivitäten eng verknüpft ist mit der Anzahl oder der Position der Akteure im Raum, wurde ein möglichst schnell zu berechnendes aber aussagekräftiges Merkmal zum Finden von Personen benötigt. Die Bewegungsanalyse allein ist hier nicht ausreichend, da Personen sich oft einfach ruhig verhalten und beispielsweise einem Seminar lauschen oder am PC arbeiten.

Zu diesem Zweck wurde ein Merkmal entwickelt, welches anhand eines Hautfarbhistogramms die Anzahl der Pixel in Teilbildern schätzen kann, die hautfarben sind.

Nach [YLW97] ist bekannt, dass sich die Farbwerte menschlicher Haut in einem engen Bereich des RGB-Farbraums ballen, und dass sich die Streuung der Werte noch reduzieren lässt, wenn man sie in den chromatischen Farbraum transformiert.

Im chromatischen Farbraum (auch: rg-Farbraum) werden zwei Farbvektoren auf denselben Punkt abgebildet, wenn sie trotz unterschiedlicher Helligkeit denselben Farbton haben. Der chromatische Farbraum entsteht aus dem RGB-Farbraum durch eine Transformation der Form

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}.$$

Um die Verteilung der Hautfarbwerte im rg-Farbraum zu charakterisieren, ist es erforder-

lich, ein Modell der Hautfarbe zu erstellen. Prinzipiell ist die Repräsentation der Hautfarbverteilung durch ein parametrisches Modell, wie z. B. eine Gaußmischverteilung, oder durch ein nicht-parametrisches Modell, z. B. ein Histogramm, möglich. In der vorliegenden Arbeit kommt letzteres zur Anwendung. Ein Histogramm H^+ wird mit Bildausschnitten initialisiert, in denen ausschließlich Hautfarbe zu sehen ist. Zum Vergleich wird ein zweites Histogramm H^- aufgebaut, welches mit Bildbereichen initialisiert wird, die keine Hautfarbe zeigen. Beide Histogramme werden normiert, so dass sich für die Summe ihrer Einträge der Wert 1 ergibt.

Die Wahrscheinlichkeiten $P(x|M^+)$ und $P(x|M^-)$ lassen sich direkt aus den normierten Histogrammen H^+ und H^- ablesen. Nach [N03] ergibt sich folgende Formel für die positive Klassifikation von x :

$$\frac{H^+(x)}{H^-(x)} > C \quad (4.1)$$

Für die Generierung des a-priori Hautfarbhistogramms stand eine Bilddatenbank der ECU ([ECU]) zur Verfügung. Diese enthält über 2000 Beispielbilder, in denen jeweils die Regionen markiert sind, die hautfarben sind (siehe Abbildung 4.4).



Abbildung 4.4: Markierte Hautfarbbereiche in Beispielbildern

Die Berechnung des Hautfarbanteils in einem Teilbild B basiert auf der pixelweisen Klassifikation nach Formel 4.1. Ist das Ergebnis größer als ein Schwellwert C , so wird das Pixel mit 1 (*hautfarben*) bewertet, ansonsten mit 0 (*nicht hautfarben*). In den Experimenten wurde ein Schwellwert von $C = 40$ verwendet. Eine Visualisierung des Merkmals zeigt

Abbildung 4.5.

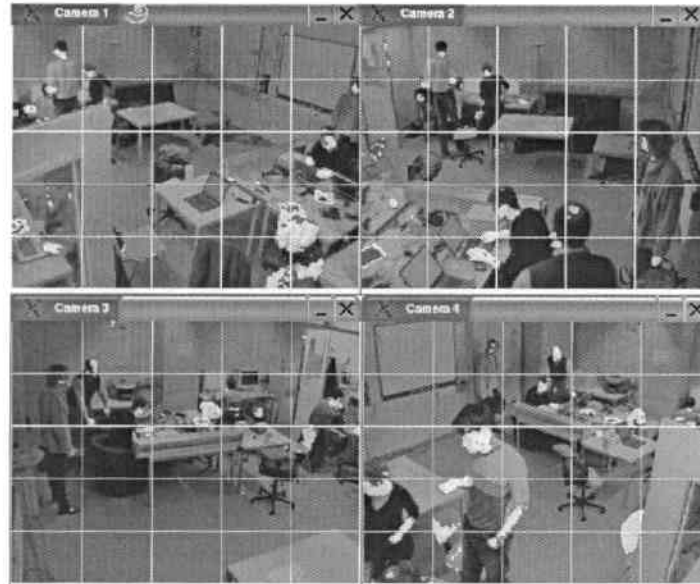


Abbildung 4.5: Hautfarbklassifikation

Das Merkmal $f_{\text{Haut}} : \mathcal{B} \rightarrow [0 \dots 1] \subset \mathbb{R}$ berechnet sich dann im Falle des Teilbildes B mit

$$f_{\text{Haut}}(B) = \frac{k_x \cdot k_y}{r_x \cdot r_y} \sum_{(x,y) \in B} \text{haut}_B(x,y)$$

wobei haut_B die beschriebene Klassifikation in B vornimmt. Der berechnete Wert spiegelt den Anteil der hautfarbenen Pixel im Teilbild B wieder. Die Konstanten $r_{x/y}$ und $k_{x/y}$ stehen wieder für die Auflösung des Gesamtbildes und die Anzahl der Teilbilder in x- und y-Richtung.

4.2.5 Optischer Fluss

Als weiteres interessantes Merkmal wurde der optischen Fluss in den Bildern in Betracht gezogen. Die Experimente beschränken sich allerdings auf die Betrachtung der Differenzbilder, da diese schneller und einfacher zu berechnen sind und auch ein gewisses Maß für die Bewegung im Bild bieten.

Bei der Berechnung des optischen Flusses werden korrespondierende Punktpaare in aufeinander folgenden Bildpaaren extrahiert. Er ist definiert durch die Verschiebungsvektoren

(u, v) , die für Bildpaare die zueinander korrespondierenden Bildpunkte aufeinander abbilden.

Hierbei wird das Helligkeitsprofil der Bilder analysiert und basierend auf der Annahme, dass sich bewegende Objekte jeweils konstante Profile haben, die Bewegungsvektoren berechnet.

Eine genaue, formale Herleitung liefert [J]. Eine Visualisierung bietet Abbildung 4.6.

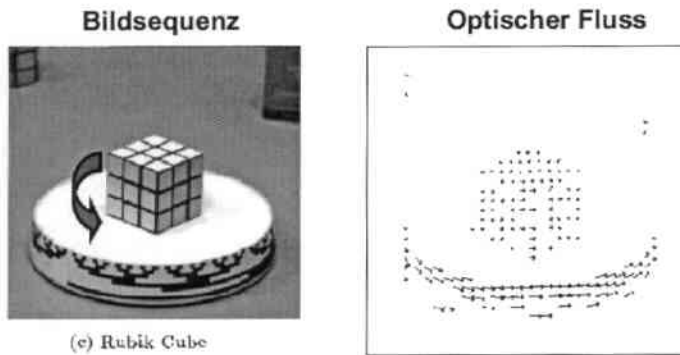


Abbildung 4.6: Bewegungsvektoren des optischen Flusses, entnommen aus [G03]

4.2.6 Intensität

Als letztes Merkmal wurde die Helligkeitsintensität f_{Int} in Teilbild B in Betracht gezogen, da eine Korrespondenz zwischen der Beleuchtungsstruktur im Raum und den dort stattfindenden Aktivitäten bestehen könnte. Die Funktion spiegelt die normalisierte Summe der Farbwerte wieder und berechnet sich wie folgt:

$$f_{\text{Int}}(B) = \frac{k_x \cdot k_y}{r_x \cdot r_y \cdot \sqrt{3 \cdot 255^2}} \sum_{(x,y) \in B} \|B(x,y)\|$$

Hierbei stehen die Konstanten $r_{x/y}$ und $k_{x/y}$ wie gehabt für die Auflösung des Gesamtbildes und die Anzahl der Teilbilder in x- und y-Richtung.

4.3 Klassifikation

Die bei der Merkmals-Auswertung entstandenen Daten spiegeln die Sichtweise wieder, die ein späterer Klassifikator auf die zu klassifizierende Situation hat. Die Daten werden auch

Eingabemuster genannt und dienen direkt als Eingabevektoren für den Klassifikator.

4.3.1 Zusammensetzung der Eingabevektoren

Angenommen, es stehen eine Menge von h aufeinander folgenden und jeweils synchronen Bildern der n Kameras zur Verfügung:

$$(b_1^1, \dots, b_1^n), \dots, (b_h^1, \dots, b_h^n)$$

Zur Auswertung wird jedes der Bilder in k_x Teile in x- und k_y Teile in y-Richtung aufgeteilt. Außerdem stehen l verschiedene Merkmale zur Verfügung, die als Eingabe eine Historie von h Frames erlauben:

$$f_1, \dots, f_l : \mathcal{B}^h \rightarrow [0, \dots, 1] \subset \mathbb{R}$$

Die Eingabevektoren setzen sich dann wie folgt zusammen:

In jeder Teilbildsequenz b_1^i, \dots, b_h^i ($i = 1 \dots n$) wird jedes Merkmal f_j ($j = 1 \dots l$) in allen Teilbildern ausgewertet. Es entsteht der Vektor

$$F_{i,j} = \begin{pmatrix} f_j(B^{1,1}(b_1^i), \dots, B^{1,1}(b_h^i)) \\ \vdots \\ f_j(B^{k_x, k_y}(b_1^i), \dots, B^{k_x, k_y}(b_h^i)) \end{pmatrix} \quad (4.2)$$

Ein kompletter Eingabevektor setzt sich dann wie folgt aus den Daten aller Merkmale in allen Teilbildsequenzen zusammen:

$$\text{Eingabevektor } v = (F_{1,1} \cdots F_{1,l} \cdots \cdots F_{n,1} \cdots F_{n,l})^\top \quad (4.3)$$

4.3.2 Auswahl und Aufbau eines Klassifikators

Für die Klassifikation der entstandenen Eingabemuster wurden künstliche neuronale Netze (KNN¹) verwendet. Diese haben sich schon in einer Vielzahl von Klassifikationsproblemen bewährt und erweisen sich auch für dieses Problem als gut geeignet.

Neuronale Netze bestehen aus einer mehrschichtigen Anordnung von Perzeptronen. Sie können anhand von vorgegebenen Eingabe-/Ausgabevektoren (überwachtes Lernen) in der so genannten Trainingsphase eine Funktion anlernen, die zu gegebenen Eingabedaten passende Ausgabedaten berechnet.

Der hier verwendete Ansatz verwendet für jede der zu erkennenden Klassen ein neuronales Netz, welches Eingabevektoren der in Gleichung 4.3 beschriebenen Form verarbeitet

¹engl.: artificial neural networks, ANN

und eine a-posteriori Wahrscheinlichkeit ausgibt, mit welcher die zu erkennende Klasse gerade aktiv ist. Der Aufbau eines solchen Netzes ist in Abbildung 4.7 illustriert.

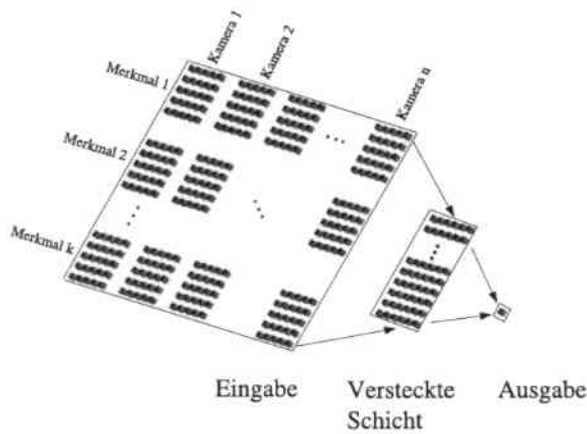


Abbildung 4.7: Die Netzarchitektur

In der Abbildung wird von einer Aufteilung der Einzelbilder in je 5×5 Teilbilder ausgegangen. Die so gruppierten Eingabeneuronen in der Eingabeschicht stehen also jeweils für die Komponenten eines Vektors $F_{i,j}$, wie er in Gleichung 4.2 beschrieben ist.

Die Größe der versteckten Schicht kann im Prinzip beliebig variiert werden. In den Experimenten wurden Netze mit 20, 40, 60 und 80 versteckten Neuronen trainiert.

Die Ausgabeschicht besteht nur aus einem einzelnen Neuron, welches eine reelle Zahl zwischen 0 und 1 ausgibt. Ist das Netz für die Erkennung der Klasse c zuständig, so nähert das Ausgabeneuron die a-posteriori Wahrscheinlichkeit $P(c|v)$ der Aktivität der Klasse c bei der Beobachtung v an.

Es wäre denkbar, als Eingabe eines Netzes einen ganzen zeitlichen Verlauf von Eingabevektoren zu verwenden, um so die Geschehnisse im Raum besser erfassen zu können. Da dies aber die Dimensionalität des Klassifizierungsproblems in die Höhe schnellen ließe, wurde hier die zeitliche Dimension nur durch die Wahl geeigneter Merkmale behandelt.

4.3.3 Weiterverarbeitung der Ausgaben

Der gewählte Ansatz liefert bei der Betrachtung von m verschiedenen Klassen c_1, \dots, c_m eine Menge K von Klassen, die zu dem betrachteten Zeitpunkt gleichzeitig aktiv sind:

$$K = \{ c_i \mid P(c_i|v) > 0,5 \}$$

Schließen sich die betrachteten Klassen gegenseitig aus, so kann alternativ auch die Klasse mit der maximalen Wahrscheinlichkeit als Klassifikationsergebnis gewählt werden. Die momentan aktive Klasse c_{aktiv} bestimmt sich dann mit

$$c_{\text{aktiv}} = \operatorname{argmax}_c P(c|v).$$

Auch eine kombinierte Anwendung der beiden Interpretationsmöglichkeiten ist denkbar.

4.3.4 Alternative Klassifikatoren

KNN sind nicht die einzige Möglichkeit, einen derartigen Erkenner zu bauen. Denkbar wäre auch die Anwendung eines Clustering-Algorithmus wie z.B. K-Means oder das Trainieren eines Bayes-Klassifikators. Die Entscheidung viel zugunsten der neuronalen Netze. Dieser diskriminative Klassifikationsansatz hat gegenüber beispielsweise Bayes-Klassifikatoren den Vorteil, dass keine klassenbedingte Wahrscheinlichkeitsverteilung angelernt werden muss. Bei einer derartig hohen Dimensionen des Merkmalsraumes sowie nur begrenztem Trainingsmaterial schien das direkte Lernen der Entscheidungsgrenzen geeigneter.

Unüberwachte Clustering-Verfahren wie K-Means haben den Nachteil, dass keine direkten Aussagen über die Semantik der gefundenen Cluster getroffen werden kann. Voraussetzung für ein solches Verfahren wäre auch, dass sich jede der semantischen Klassen auf ein Ballungsgebiet im Merkmalsraum beschränkt. Eine Untersuchung der Korrelation von gefundenen Clustern mit den Trainingsdaten wäre dennoch ein interessantes Experiment für die Zukunft.

Kapitel 5

Experimente

Der in Kapitel 4 beschriebene Ansatz wurde implementiert und seine Erkennungsleistung bezüglich der in Sektion 3.3 diskutierten, so genannten "globalen Zustände"¹ evaluiert.

5.1 Datenbasis

In unseren Experimenten standen uns ca. 149 Stunden Videomaterial zur Verfügung. Wir haben die gesamte Datenbasis aufgeteilt in zwei Teile, einen zum Trainieren und einen zum Testen der Klassifikatoren. Das Trainingsset enthält 92 Stunden Videomaterial, welches im Laufe des Jahres 2003 sowie im Frühjahr 2004 aufgenommen wurde. Getestet wurde auf 57 Stunden umfassenden Daten, die im Sommer 2004 aufgenommen wurden. In dem Zeitraum zwischen den Aufnahmen der Trainings- und Testdaten wurden aus anderweitigen Gründen die Kameras um jeweils ca. ein Viertel ihres Sichtfeldes gedreht. Das Klassifikationsproblem auf den Testdaten wurde somit erschwert.

Wir haben für jede der zu erkennenden Klassen eine zufällige Auswahl von positiven und negativen Trainingsbeispielen getroffen. Es wurden jeweils für das Training und das Testen ca. 2500 Stichproben aus der Datenbasis entnommen. Die Stichproben liegen mindestens 20 Sekunden, in der Regel jedoch noch weiter auseinander. Eine ausreichende Streuung sollte so gewährleistet sein.

5.2 Training der Klassifikatoren

Zum Trainieren der neuronalen Netze wurden diese mit zufälligen Gewichten initialisiert. Die Trainingsdaten wurden aufgeteilt in 20 Prozent Cross-Evaluationsdaten und 80 Prozent eigentliche Trainingsdaten. Im so genannten "Batch-Learning"-Verfahren wurden dem jeweiligen Netz dann alle zur Verfügung stehenden Trainingsdaten präsentiert und die

¹"Seminar findet statt", "Laboralltag", "Sonstiges"

Gewichte entsprechend des Fehlers aktualisiert. Dieser Vorgang wurde bis zu 100 mal wiederholt. In jedem Durchgang wurde auch der mittlere quadratische Fehler (MSE²) auf den Cross-Evaluationsdaten ermittelt. Es wurde am Ende dasjenige Netz gewählt, dass auf den Cross-Evaluationsdaten die besten Werte erzielt hat. Die Entwicklungen des MSE auf den Trainingsdaten (gestrichelte Linie) und den Cross-Evaluationsdaten (durchgezogene Linie) sind in den Abbildungen 5.1 und 5.2 beispielhaft grafisch dargestellt. Alle später aufgeführten Ergebnisse sind gemittelt über alle möglichen Variationen der Wahl des 20-prozentigen Cross-Evaluationsanteils ("Round Robin").

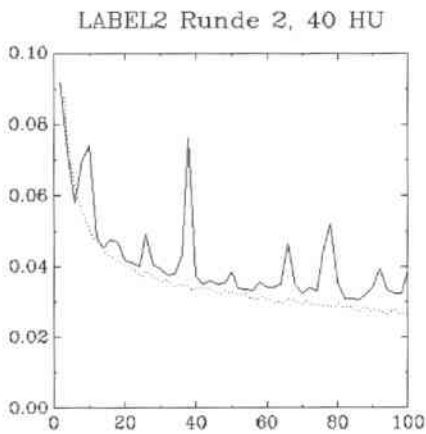


Abbildung 5.1: Entwicklung MSE Beispiel 1

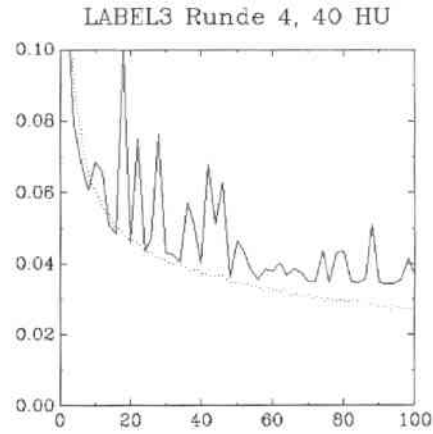


Abbildung 5.2: Entwicklung MSE Beispiel 2

5.3 Ergebnisse

Im Folgenden sind die Ergebnisse der Experimente genauer aufgeführt. Es werden jeweils die Erkennungsrate insgesamt sowie die Konfusionsmatrizen angegeben. Es fanden sowohl Experimente statt, in denen die Bilder als ganzes betrachtet wurden, als auch Experimente, in denen die Bilder in 16 bzw. 25 Teilbilder aufgeteilt wurden.

5.3.1 Ergebnisse ohne Betrachtung von Teilbildern

Schon bei einer Betrachtung der Bilder als ganzes hat die Klassifikation eine gute Erkennungsleistung geliefert. Als bestes Merkmal hat sich die Hautfarbe herausgestellt. Durch Hinzunahme weiterer Merkmale haben sich die Ergebnisse verschlechtert. Die beste Erkennungsleistung wurde mit 5 versteckten Units erreicht. Die Gesamterkennungsleistung lag

²engl. für "mean square error"

bei 88.6%. Genauere Ergebnisse finden sich in der Konfusionsmatrix in Tabelle 5.1.

Ziel \ Ergebnis	Sonstiges	Alltag	Seminar	Datenanteil
Sonstiges	60.2%	39.8%	0.0%	11.5%
Alltag	0.1%	81.9%	18.0%	37.8%
Seminar	0.0%	0.0%	100.0%	50.7%

Tabelle 5.1: Ergebnisse bei folgendem Experiment: keine Regionen, Hautfarbe, 5 versteckte Units. Gesamterkennungsrate: 88.6%

Als zweitbeste Merkmalskombination hat sich Hautfarbe und Differenzbild erwiesen. Hier haben Netze mit 10 versteckten Units das beste Ergebnis mit einer Gesamterkennungsrate von 88.0% geliefert (siehe Tabelle 5.2).

Ziel \ Ergebnis	Sonstiges	Alltag	Seminar	Datenanteil
Sonstiges	61.1%	38.9%	0.0%	11.5%
Alltag	3.0%	80.2%	16.8%	37.8%
Seminar	0.0%	0.0%	100.0%	50.7%

Tabelle 5.2: Ergebnisse bei folgendem Experiment: keine Regionen, Hautfarbe und Differenz, 10 versteckte Units. Gesamterkennungsrate: 88.0%

5.3.2 Ergebnisse mit 4x4 Teilbildern

Bei einer Aufteilung der Bilder in 4x4 Teilbilder haben sich folgende Ergebnisse ergeben. Am besten funktioniert hat wieder das Merkmal Hautfarbe. 40 versteckte Units haben hier die beste Erkennungsleistung geliefert. Die entstandene Konfusionsmatrix ist in Tabelle 5.3 wiederzufinden. Die Erkennungsleistung insgesamt betrug in dieser Konfiguration 84.1%.

Ziel \ Ergebnis	Sonstiges	Alltag	Seminar	Datenanteil
Sonstiges	46.5%	53.2%	0.2%	11.5%
Alltag	21.0%	74.2%	4.8%	37.8%
Seminar	0.0%	0.0%	100.0%	50.7%

Tabelle 5.3: Ergebnisse bei folgendem Experiment: 4x4 Regionen, Hautfarbe, 40 versteckte Units. Gesamterkennungsrate: 84.1%

Zweitbestes Merkmal war hier das Differenzbild. Mit nur zwei versteckten Units wurde hier eine Erkennungsleistung von 76.0% erzielt (siehe Tabelle 5.4).

<i>Ziel</i> \ <i>Ergebnis</i>	Sonstiges	Alltag	Seminar	Datenanteil
Sonstiges	97.8%	1.2%	1.0%	11.5%
Alltag	39.0%	37.2%	23.9%	37.8%
Seminar	0.0%	0.0%	100.0%	50.7%

Tabelle 5.4: Ergebnisse bei folgendem Experiment: 4x4 Regionen, Differenz, 2 versteckte Units. Gesamterkennungsrate: 76.0%

5.3.3 Ergebnisse mit 5x5 Teilbildern

Die besten Ergebnisse insgesamt lieferten die Klassifikatoren bei einer Bildaufteilung in 5x5 Teilbilder. Bestes Merkmal war hier wieder die Hautfarbe. Mit 80 versteckten Units wurde hier eine Erkennungsleistung von 92.6% erreicht (siehe Tabelle 5.5).

<i>Ziel</i> \ <i>Ergebnis</i>	Sonstiges	Alltag	Seminar	Datenanteil
Sonstiges	46.5%	53.3%	0.2%	11.5%
Alltag	1.3%	96.7%	2.0%	37.8%
Seminar	0.0%	0.0%	100.0%	50.7%

Tabelle 5.5: Ergebnisse bei folgendem Experiment: 5x5 Regionen, Hautfarbe, 80 versteckte Units. Gesamterkennungsrate: 92.6%

Zweitbestes Merkmal war hier die Bewegung über einen zeitlichen Verlauf. Mit 10 versteckten Units wurde eine Erkennungsleistung von 80.8% erreicht (siehe Tabelle 5.6).

<i>Ziel</i> \ <i>Ergebnis</i>	Sonstiges	Alltag	Seminar	Datenanteil
Sonstiges	16.6%	83.3%	0.1%	11.5%
Alltag	3.5%	91.7%	4.7%	37.8%
Seminar	0.0%	12.7%	87.3%	50.7%

Tabelle 5.6: Ergebnisse bei folgendem Experiment: 5x5 Regionen, Bewegung über Zeit, 10 versteckte Units. Gesamterkennungsrate: 80.8%

5.3.4 Diskussion

Wie schon weiter oben erwähnt beruhen die hier angegebenen Ergebnisse auf Testdaten, die komplett getrennt von den Trainingsdaten mindestens ein halbes Jahr später aufgenommen wurden. In dem dazwischen liegenden Zeitraum wurden die Kameras in dem Raum leider etwas gedreht (je um ca. einen viertel des Sichtfeldes).

Es fällt auf, dass die Erkennungsleistung bei einer Bildaufteilung in 16 Teilregionen schlechter ist als bei einer Betrachtung der Bilder als Ganzes. Es könnte sein, dass die Drehung

der Kameras für den einfacheren Erkennen weniger gravierend war als für die Erkennen mit Bildaufteilung. Insgesamt betrachtet liefert jedoch alle Klassifikatoren trotz der Kameradrehung recht robuste Ergebnisse. Die besten Ergebnisse wurden mit einer Bildaufteilung von 5x5 Bildern erzielt. In Zukunft wäre es interessant, mit anderen Bildaufteilungen wie beispielsweise 2x2, 3x3 oder auch 6x6 und mehr Teilbildern zu experimentieren.

Auf eine weitere Auffälligkeit stößt man bei der Betrachtung der Merkmalskombinationen. Das Merkmal Hautfarbe hat in allen Konfigurationen am besten funktioniert. Die Hinzunahme weiterer Merkmale hat sich immer negativ auf die Erkennungsleistung ausgewirkt, obwohl die hinzu genommenen Merkmale für sich betrachtet auch zu nennenswert guten Ergebnissen geführt haben. Es bleibt zu vermuten, dass dieser Effekt durch eine Vergrößerung der Datenbasis vermeidbar ist und sich in diesem Fall noch bessere Erkennungsleistungen erzielen lassen.

Kapitel 6

Zusammenfassung und Ausblick

In dieser Studienarbeit wurde ein Verfahren entwickelt, Aktivitäten im intelligenten Raum anhand von visuell erfassten Merkmalen zu klassifizieren. Neben bildverarbeitenden Elementen kamen hierbei künstliche neuronale Netze zum Einsatz. Auch eine Diskussion und Strukturierung möglicher Klassifikationsziele im Bereich der Aktivitätenerkennung sowie eine umfassende Datensammlung und -auswertung waren Teil dieser Arbeit. Der vorgestellte Ansatz wurde auf seine Erkennungsleistung bezüglich der grundlegenden Zustände im Raum hin getestet.

Die vorliegenden Experimente haben gezeigt, dass mit dem Verfahren eine recht robuste Unterscheidung der Zustände "Seminarveranstaltung", "Laboralltag" und "Sonstiges" möglich ist. Über die Erkennungsleistung bezüglich kürzer andauernder oder sich auf eine Örtlichkeit beschränkender Aktivitäten kann jetzt noch keine Aussage gemacht werden. Grundsätzlich ist der entwickelte Klassifikator jedoch für die Erkennung derartige Aktivitäten geeignet. Aufgrund der aufwendigen Navigation in großen Videosequenzen und dem dementsprechend hohen Aufwand für das Labeln der Daten musste in dieser Arbeit auf derartige Evaluationen verzichtet werden.

Da die verwendeten Merkmale nicht speziell auf das Szenario zugeschnitten sind, wäre es interessant zu prüfen, ob die verwendete Technik auch in anderen Bereichen, in denen Aktivitäten jeglicher Art erkannt werden sollen, ähnlich gute Ergebnisse liefert. Auch eine Erweiterung des Merkmalkatalogs um weitere visuelle oder auch auditive Merkmale ist denkbar.

Abbildungsverzeichnis

3.1	Beispielaufnahme der vier Kameras	10
4.1	Aufbau eines Einzelbildes	15
4.2	Differenzbild	16
4.3	Bewegungsabläufe über die Zeit	18
4.4	Markierte Hautfarbbereiche in Beispielbildern	19
4.5	Hautfarbklassifikation	20
4.6	Bewegungsvektoren des optischen Flusses, entnommen aus [G03]	21
4.7	Die Netzarchitektur	23
5.1	Entwicklung MSE Beispiel 1	26
5.2	Entwicklung MSE Beispiel 2	26

Literaturverzeichnis

- [B97] M. BRAND:
Learning concise models of human activity from ambient video via a structure-inducing M-step estimator, *1997*
- [BK00] M. BRAND, V. KETTNAKER:
Discovery and segmentation of activities in video, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000
- [BOP96] M. BRAND, N. OLIVER, A. PENTLAND:
Coupled Hidden Markov Models for complex action recognition, *Proc. of CVPR97*, S. 994-999, 1996
- [CP00] BRIAN CLARKSON AND ALEX PENTLAND:
Framing through peripheral perception, *ICIP 2000*
- [CP99] CLARKSON, B. AND A. PENTLAND:
Unsupervised clustering of ambulatory audio and video, *ICASSP'99*, 1999
- [ECU] S. L. PHUNG: Visual Information Processing Research Group, School of Engineering and Mathematics Edith Cowan University, Western Australia, <http://www.cs.ecu.edu/>
- [G03] M. GIESE:
Lernmethoden in Computervision und Computer Grafik, *Januar 2003*
- [GCBBB03] DANIEL GATICA-PEREZ, IAIN MCCOWAN, MARK BARNARD, SAMY BENGIO, HERVE BOURLARD:
On automatic annotation of meeting databases, *IDIAP-RR 03-06*
- [GJH01] A. GALATA, N. JOHNSON, D. HOGG:
Learning variable length markov models of behaviour, *International Journal on Computer Vision, IJCV*, S. 398-413, 2001
- [J] B. Jähne: Digitale Bildverarbeitung, *Springer Verlag*
- [LAL01] K. VAN LAERHOVEN, K. A. AIDOO, S. LOWETTE:
Real-time analysis of data from many sensors with neural networks, *IEEE Press*, 2001

- [N03] KAI NICKEL:
Erkennung von Zeigegesten basierend auf 3D-Tracking von Kopf und Händen, *Diplomarbeit, Karlsruhe, 2003*
- [NIST] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY:
NIST Smartflow, <http://www.nist.gov>
- [OHG] NURIA OLIVER, ERIC HORVITZ, ASHUTOSH GARG:
Layered representations for human activity recognition
- [PJ] M. PRICE, GERHARD DE JAGER:
An unsupervised clustering approach to location classification, *University of Cape Town, Department of Electrical Engineering*
- [PN94] R. POLANA, R. NELSON:
Low level recognition of human motion, *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects, Austin, S. 77-82, 1994*
- [TSKO94] K. TAKAHASHI, S. SEKI, H. KOJIMA, R. OKA:
Recognition of dexterous manipulations from time-varying images, *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects, Austin, S. 23-28, 1994*
- [WB98] A. WILSON, A. BOBICK:
Recognition and interpretation of parametric gesture, *Proc. of International Conference on Computer Vision, ICCV'98, S. 329-336, 1998*
- [YOI92] J. YAMATO, J. OHYA, K. ISHII:
Recognizing human action in time-sequential images using Hidden Markov Model, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition, S. 379-385, 1992*