ao. Univ.-Prof. Dr. Horst Eidenberger
Allgemein beeideter und gerichtlich zertifizierter Sachverständiger für Informatik
Institut für Softwaretechnik und Interaktive Systeme der TU Wien

Neugierige Computer*

Zum Stand der automatisierten Massenüberwachung

Anonymität ist Vergangenheit: Der Leistungsfähigkeit moderner Computer und der Findigkeit unserer Wissenschaftler ist es zu verdanken, dass die Staaten der Ersten Welt heute veritable Überwachungsgesellschaften sein können. Stimmt das?

Im Folgenden sehen wir uns an, welche Methoden aktuell zur Massenüberwachung eingesetzt werden. Dabei fokussieren wir auf die derzeit leistungsfähigsten Ansätze und versuchen in allgemein verständlicher Sprache, ihre Funktionsweise aus ähnlichen natürlichen Systemen, insbesondere der menschlichen Wahrnehmung herzuleiten.

1. Massenüberwachung: Was ist das?

Per definitionem handelt es sich bei Massenüberwachung um die flächendeckende Beobachtung ganzer Bevölkerungen. Die Beobachtung muss weder mit der Zustimmung der Beobachteten erfolgen noch zu deren Vorteil sein. Wesentlich ist breitbandige Datenerfassung, die – bei den heutigen Personalkosten – den Einsatz maschineller Überwachungsmethoden bedingt. Klassische Anwendungen sind *closed circuit television* (CCTV), das heißt Videoüberwachung, das Abhören von Telefonen (wo etwa die Bundesrepublik Deutschland 20 Jahre nach dem Ende der DDR wieder weltweit führend ist) und *network sniffing*, das Abhören des Datenverkehrs im Internet. Wir wollen uns im Weiteren auf die audiovisuellen Medien konzentrieren, weil dort die eingesetzten Erkennungsmethoden am reizvollsten sind.

Ziel der Massenüberwachung ist zumeist das Erkennen und Verfolgen relevanter Individuen und Situationen. Was relevant ist, definiert der Überwacher, zumeist eine öffentliche Körperschaft (Polizei, Kommunen etc). Die gewonnenen Überwachungsdaten müssen dazu gespeichert, indiziert und durchsucht werden. Zur Speicherung personenbezogener Daten gibt es seit einigen Jahren auch eine EU-Direktive, die festschreibt, welche Institution welche Daten (zB Mobiltelefongespräche) wie lange (üblicherweise sechs bis 24 Monate) aufzubewahren hat.[1] Entscheidender Schritt ist die Indizierung, bei der die Datenflut dem Überwacher effektiv zugänglich gemacht wird.

2. Schauen: Aus Sensationen Merkmale machen

Die Indizierung audiovisueller Massendaten (hunderter Kameras und Mikrofone) vollzieht sich stets in zwei Schritten: der Extraktion von Merkmalen und dem Verstehen der Bedeutung der Merkmale. Die beiden folgenden Abschnitte behandeln diese Prozesse.

Als Erstes ist es wichtig, zu verstehen, was Merkmale von Medien (englisch *features*) sind. Bei einem Merkmal handelt es sich gewissermaßen um eine statische, stark verlustbehaftete Zusammenfassung des zeitabhängigen audiovisuellen Datenstroms. Die endlosen Datenflüsse von Mikrofonen und Kameragruppen haben eine für die Auswertung ungünstige Form: Sie sind zu umfangreich, zu schnell und orientieren sich zu stark an unserer analogen Lebenswelt. Das heißt, es ist für Computerverfahren sehr schwer, aus dem Gewirr des Überwachungsvideos eines öffentlichen Platzes mit allen seinen Störungen (Sonnenstand, Regen, Fettflecken, Signalrauschen usw) logische Schlüsse zu ziehen. Gute Merkmale sind hingegen einfach. Typische Beispiele sind Farbverteilungen und Helligkeitskanten, wie sie in den ersten Stufen des menschlichen visuellen Systems wahrgenommen werden, und Rhythmus und Frequenzspektrum, wie sie die Gehörschnecke des Menschen herausfiltert. Diesen Merkmalen ist gemein, dass sie digital sind. Sie können in endlichen Zahlenkolonnen (Vektoren) dargestellt werden, die sich hervorragend für die Weiterverarbeitung durch auf mathematischen Modellen basierende Logik-Algorithmen eignen.

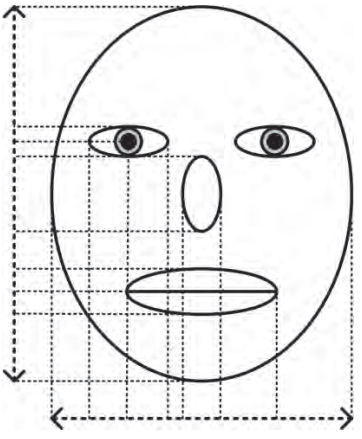

Abbildung 1: Gesichter liefern zahlreiche biometrische Merkmale

* Abdruck eines populärwissenschaftlichen Zeitschriftenartikels des Autors, zuerst erschienen in iX – Magazin für professionelle Informationstechnik 2/2009, 122.

Abbildung 1 zeigt als Beispiel für visuelle Merkmale die Vielzahl von Werten, die aus der Biometrie eines Gesichtes gewonnen werden können. Besteht erst einmal der Verdacht, dass sich in einem Videobild ein Gesicht befindet (zB aufgrund der typischen Gesichtsfarben), können schnell Augen (rund, schattig) und Nasenspitze (fast immer besser ausgeleuchtet als ihre Umgebung) sowie in der Folge der Mund gefunden werden. Die Verhältnisse von Lippenhöhe und -breite sowie von Augenabstand und Augen-Mund-Abstand liefern passable Merkmale für die Identifikation von Personen.[2]

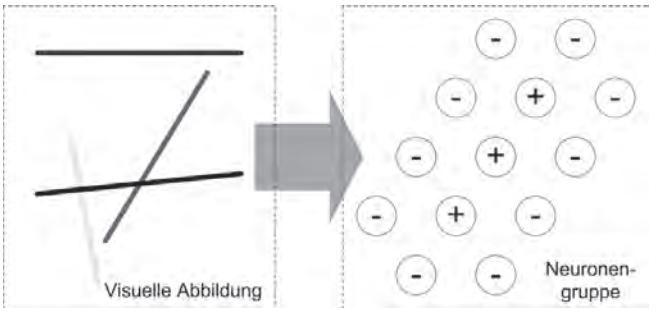

Abbildung 2: Das Gehirn enthält für jeden Kantentyp eigene Neuronengruppen

Die einzelnen Bezugspunkte des Gesichtes (aber auch jedes anderen Objektes in einem visuellen Medium) werden mittels Segmentierung und Kantenextraktion gewonnen. Die Vielzahl der existierenden Ansätze folgt dabei einem Verfahren, das auch im menschlichen Gehirn neuronal implementiert ist. Abbildung 2 zeigt das Prinzip. Einzelne Neuronengruppen reagieren auf Hell-Dunkel-Kanten (zB die Umrisse der Augen in Abbildung 1), indem ihre Netzhautdetektoren entsprechend auf die auftreffenden Photonen reagieren und verstärkende und abschwächende Synapsen gebildet werden. Die künstlichen Kantenerkennungs-Algorithmen basieren freilich zumeist auf algorithmisch effizienteren Matrix-Operatoren, die den Videodatenstrom räumlich falten.

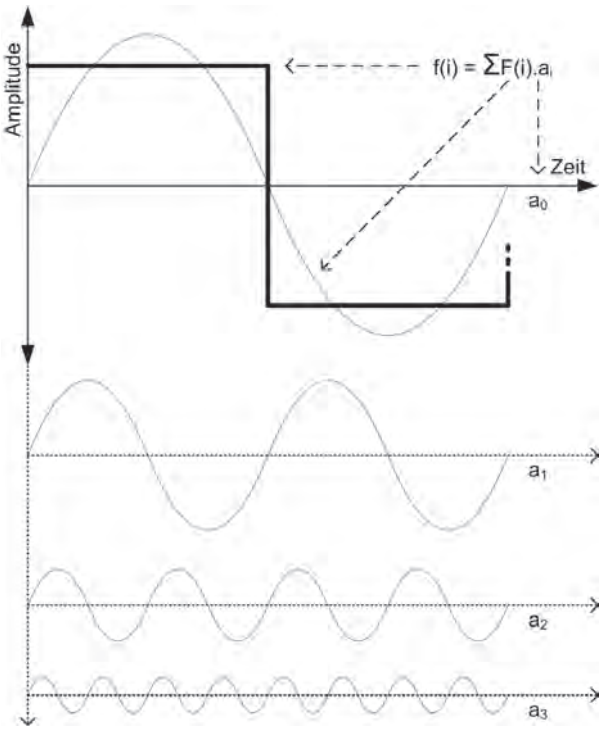

Abbildung 3: Die Fourier-Transformation stellt Signale als Summe von Sinusschwingungen dar

Wo mit Audiodaten überwacht wird, wird zunächst meist ein sogenanntes Spektrum erzeugt. Abbildung 3 skizziert den – sehr mathematischen – Prozess, der die Funktionsweise der Gehörschnecke des Menschen nachahmt. Beim Hören wird der ein-dimensionale Eingabedatenstrom in eine Folge von Grundfrequenzen aufgeteilt. Zumeist finden dafür die Sinusschwingungen der Fourier-Transformation Anwendung. Das zugrunde liegende Modell der Integraltransformationen nach Laplace sorgt dafür, dass die Summe einzelner Schwingungen mit unterschiedlicher Frequenz letztlich wieder dem Original-Signal entspricht. Aus den Sinusschwingungen (die Merkmale sind ihre Koeffizienten) lassen sich charakteristische Aussagen über den Inhalt des Audio-Datenstroms ableiten.

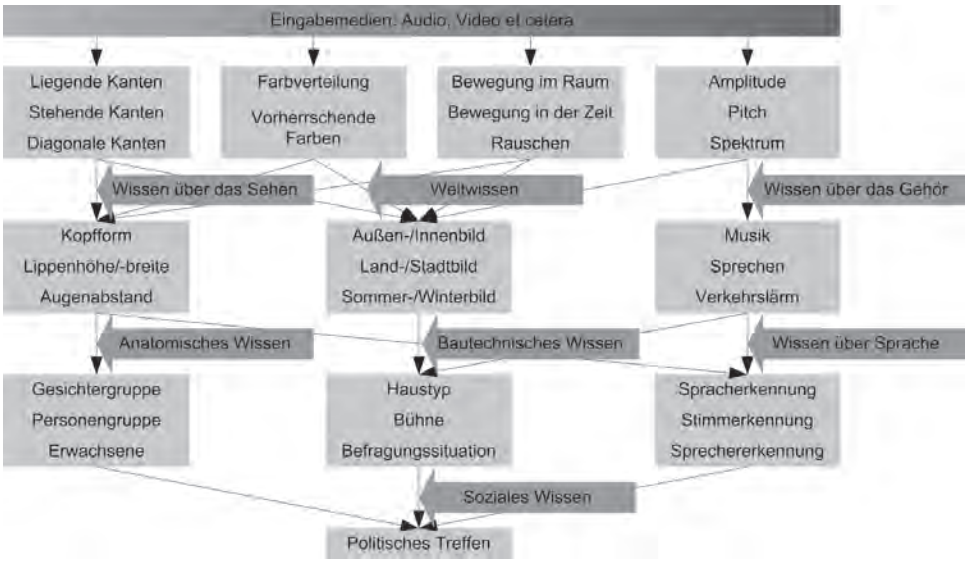

Abbildung 4: Komplexere Merkmale basieren auf einfacheren und Zusatzwissen

Diese Beispiele zeigen bereits das Hauptproblem der Merkmalsextraktion. Die gewonnenen Merkmale sind allzu primitiv. Was sagen schon die Statistiken über Anzahl und Lage von Hell-Dunkel-Kanten und über enthaltene Grundfrequenzen aus? Um diesem Übel beizukommen, bedient man sich der Methoden der künstlichen Intelligenz zur semantischen Anreicherung. Abbildung 4 skizziert den Prozess. Durch die Einbringung semantischen Wissens (zB über die Organi-

sation von Gesichtern, den Aufbau von Sprache, die Architektur von Häusern) und logisches Schließen werden aus den einfachen Merkmalen komplexere, aussagekräftigere gewonnen. Wiederholtes Anwenden dieses Prozesses erlaubt inhaltlich relevantere Schlussfolgerungen (leider bei – teilweise stark – abnehmender Zuverlässigkeit der Aussagen).

3. Sehen: Das Geschaute verstehen

Wie auch immer die Merkmale gewonnen und angereichert wurden, nach der Extraktion liegen sie als Datenvektoren vor. Das ist ein wichtiges Prinzip. Es bedeutet, dass jeder Teileigenschaft des Merkmalsbündels eines Objektes (zB eines Bildes einer Überwachungskamera) eine Bedeutung zugeordnet wird (etwa „Augenabstand“) und dass sich dieselben Bedeutungen zweier Objekte an denselben Stellen ihrer Merkmalsbündel (Vektoren) befinden.

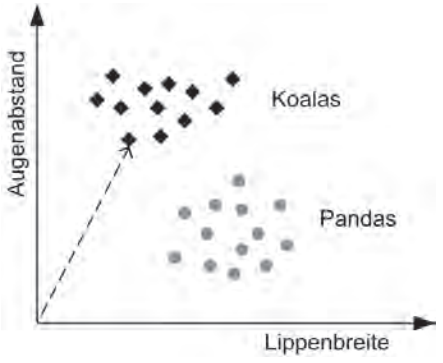

Abbildung 5: Merkmalsräume können viel-dimensional sein

Abbildung 5 veranschaulicht die zwei-dimensionalen Merkmalsvektoren zweier Populationen von Tieren, wie sie zB eine Überwachungskamera in einem Zoo aufnehmen könnte. Wäre es die Aufgabe des Überwachungssystems, ständig zu prüfen, dass sich Koalas und Pandas nicht buchstäblich ins Gehege kommen, so würden die durch Merkmalsextraktion gewonnenen Datenvektoren nach ihren typischen Eigenschaften in zwei Gruppen geschieden. Dabei wird grundsätzlich jeder Merkmalsvektor als Punkt in einem metrischen Vektorraum interpretiert (was mathematisch einigermaßen abenteuerlich sein kann), wodurch es möglich wird, Abstände zu messen. Basierend auf dieser Einteilung könnte nun verfolgt werden, wo sich die Tiere (identifiziert anhand ihrer Merkmalsvektoren) aufhalten.

Diese sogenannte **Klassifikation** ist eine typische Anwendung der Massenüberwachung. Dabei wird ein Objekt (Merkmalsvektor) einer von mehreren Gruppe zugeordnet. Klassifikation kann dort erfolgreich eingesetzt werden, wo große Gruppen (Populationen) voneinander unterschieden werden sollen: zB bei Tiere im Zoo oder Fangruppen im Stadium – gleiche Technik bei gleichem Sozialverhalten. Sie versagt, wo etwas erkannt werden soll, das bis zu einem gewissen Grad neu ist. Typisches Beispiel dafür ist das Gesicht einer bestimmten Person, die gefunden werden soll. Dann sind die Suchmethoden des *information retrieval* gefragt, die zu einem Merkmalsvektor die ähnlichsten einer Population finden und diese nach Relevanz reihen. Diese **Ähnlichkeitssuche** wird meist durch – zB euklidische – Abstandsmessung implementiert, wobei größere Distanz als größere Unähnlichkeit interpretiert wird. Im Gegensatz dazu werden zur Klassifikation meist Methoden des Maschinenlernens oder der Stochastik eingesetzt.

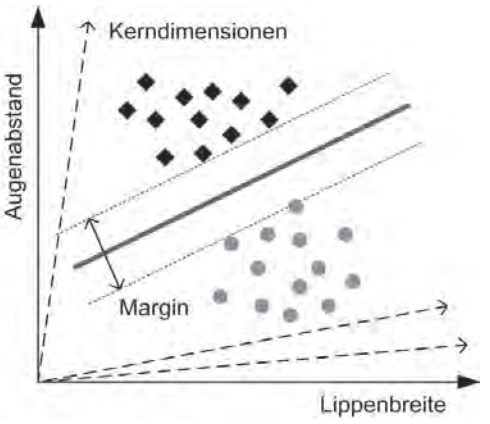

Abbildung 6: Die Support-Vector-Maschine zerlegt in Gut und Böse

Abbildung 6 zeigt eine der derzeit leistungsfähigsten Methoden des Maschinenlernens:[3] die Klassifikation mithilfe einer so genannten Support-Vector-Maschine (SVM). Die SVM beruht auf einer Reihe von Prinzipien, die für das Maschinenlernen typisch sind. Zunächst folgt sie der oben beschriebenen Grundannahme, Merkmalsvektoren als Punkte in einem Vektorraum zu sehen. In diesem Vektorraum versucht sie, lediglich zwei Gruppen (sozusagen die „Guten“ von den „Bösen“) zu unterscheiden, indem sie eine Grenze zwischen ihnen zieht (englisch *margin*). Da der Merkmalsraum im Allgemeinen mehr als zwei Dimensionen hat, wird es sich dabei um eine Hyperebene handeln. Das nächste wichtige Prinzip ist, dass sie diese Grenze mithilfe menschlichen Zusatzwissens zu erlernen versucht. Dazu muss man der SVM eine Stichprobe von markierten Punkten bereitstellen, die jeweils eindeutig einer der beiden Gruppen zugeordnet sind. Schließlich erfolgt das Lernen nicht im Merkmalsraum, sondern einem daraus erzeugten Raum, der noch viel mehr Dimensionen als dieser hat. Das hat einen einfachen Grund: Je höher-dimensional der untersuchte Vektorraum ist, umso größer werden die Abstände zwischen der konstant bleibenden Anzahl von Punkten der Stichprobe. Dadurch wird es leichter, die Grenze zu ziehen. Diesen Ansatz nennt man **kernel-basiertes Lernen**. Als Kernel bezeichnet man die Funktion zur Transformation der Datenpunkte in den höher-dimensionalen Raum.

Einer der derzeit leistungsfähigsten Vertreter der stochastischen Ansätze sind die sogenannten Mixtur-Modelle. Bei ihrem häufigsten Vertreter, dem Gauß’schen Mixtur-Modell (GMM), wird die Normalverteilung verwendet, um Punktwolken als Gruppen zu erkennen. Abbildung 7 zeigt ein Beispiel. Im Gegensatz zum Maschinenlernen wird keine Stichprobe benötigt, aber man muss dem Algorithmus einen Tipp geben, um wie viele Gruppen es sich handeln

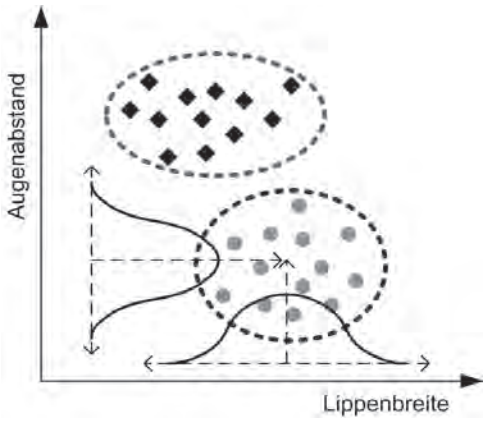

Abbildung 7: Mixtur-Modelle bauen Zäune um Merkmalswolken

wird. Dann versucht die GMM, die Gruppen als Punktwolken einzufangen, die in jeder Dimension n-fach normalverteilt sind. Dazu wird ein zweistufiger Algorithmus iteriert. Im ersten Schritt werden die Parameter der einzelnen Verteilungen geschätzt. Im zweiten wird die Schätzung mit dem Merkmalsraum verglichen. Anhand der Unterschiede werden die Parameter so lange angepasst, bis die Gruppen ausreichend gut unterschieden werden können. Wesentlich an diesem Prozess ist, dass die Annahme der Normalverteilung in den Daten tatsächlich gegeben ist. Abbildung 7 zeigt zwei sehr schön separierte Punktwolken. Wären sie ineinander verschmiert, würde die Unterscheidung dem GMM wesentlich schwerer fallen oder überhaupt scheitern.

Neben diesen beiden gibt es noch eine Vielzahl anderer Modelle. Sie kommen aus den genannten Disziplinen ebenso wie aus dem *operations research*, der künstlichen Intelligenz und anderen. Wichtige Ansätze sind zB Markov-Modelle zur Spracherkennung (stochastisch, aber mit Stichprobenlernen) und Bayes-Klassifikatoren zur schnellen Klassifikation.

4. Was ist möglich? Was realistisch?

Trotz aller Komplexität der eingesetzten Methoden bleibt die automatische Massenerkennung im Einzelnen immer noch weit hinter dem menschlichen Erkennungsapparat zurück. Neben der allgemein geringeren Qualität der gelieferten Aussagen ist insbesondere das Problem der *false positives* relevant. So nennt man die fehlerhafte Zuordnung eines stattfindenden Ereignisses (zB einer Einzahlung am Schalter) mit einem gesuchten (einem Banküberfall). Formal betrachtet muss ein maschinelles Verfahren, das menschliche Kognition nachahmt, im Einzelfall stets hinter dem Menschen zurückbleiben. Es kann also nicht sein, dass ein Computer zB Gesichter erkennt, die ein menschlicher Experte nicht erkennen würde. Das schließt den Einsatz algorithmischer Analysemethoden im Beweisverfahren weitgehend aus.

Andererseits ermüden Computer nicht und schwanken auch nicht in ihrem Urteil. Bei der massenhaften Auswertung umfassender Mediendatenströme (ein Netzwerk von fünf typischen CCTV-Kameras liefert jede Sekunde zirka 142 Megabyte an Daten!) sind das gewichtige Argumente. Bei solchen Datenlawinen macht die Speicherung nur Sinn, wenn maschinell indiziert wird. Sonst wäre es nur möglich, Ereignisse aufzufinden, deren genaue Raumzeitkoordinaten man kennt. Daher wächst der Markt der Massenüberwachungssysteme derzeit stark. Schon viele freie Produkte liefern gute Software für Personen-, Bewegungs- und Objekterkennung. Besonders im Trend ist dabei die Erkennung der Nummernschilder von Autos (zB durch Kantenerkennung und die Anwendung von SVM auf die Zeichen des Schildes). Was noch fehlt sind Module für die Erkennung und Beurteilung von Situationen auf semantisch höherer Ebene (Verkehrsunfälle, Risikosituationen etc). Da dazu ein Erkennungsapparat von der Leistungsfähigkeit unseres Gehirns benötigt würde, ist ein solcher Quantensprung in nächster Zeit nicht zu erwarten.

5. Fazit

Massenüberwachung ist heute allgegenwärtig. Es stellt sich daher die Frage, wo die Grenzen der Erkennbarkeit liegen und wie man sich als Individuum vor unrechtmäßigem Zugriff effizient schützt. Die wesentlichen Defizite heutiger System liegen sicherlich in der Merkmalsextraktion. Macht man wesentliche Merkmale unkenntlich, wird die Erkennung schnell unmöglich. Wer eine Sonnenbrille aufsetzt und sich einen falschen Bart anklebt, ist maschinell praktisch nicht mehr zu erkennen. Sein Abbild versinkt zwar für vielleicht zwei Jahre in einem riesigen Medienarchiv, ist dort aber so unauffindbar wie die Bundeslade im ersten Indiana-Jones-Film. Bei anderen Artefakten, die nicht durch Persönlichkeitsrechte geschützt sind (etwa Autos), ist diese Form der Camouflage wesentlich schwieriger. Die Erkennung von Nummernschildern ist heute schon sehr ausgereift. Hier können nur zivilgesellschaftliche Maßnahmen das Individuum vor dem Big Brother schützen.

Anmerkungen:

[1] Richtlinie des Europäischen Parlaments und des Rates über die Vorratsspeicherung von Daten, online abrufbar unter http://register.consilium.eu.int/pdf/de/05/st03/st03677.de05.pdf.
[2] Gesichtserkennungswebsite http://www.face-rec.org/ (zuletzt online 8. 2. 2011).
[3] Maschinenlernen Website http://www.machinelearning.org/ (zuletzt online 8. 2. 2011).

Korrespondenz:
ao. Univ.-Prof. Dr. Horst Eidenberger
Institut für Softwaretechnik und Interaktive Systeme,
TU Wien
Favoritenstrasse 9/1882, 1040 Wien
E-Mail: eidenberger@tuwien.ac.at