State-of-the-Art Report

Autoren:

Martin Köstinger

Paul Wohlhart

Peter M. Roth

Horst Bischof

Institut für Maschinelles Sehen und Darstellen,
Technische Universität Graz
{koestinger, wohlhart, pmroth, bischof}@icg.tugraz.at

Projekttitel: Multimedia Documentation Lab: Wissensrepräsentation und -Organisation bei multimedialen Inhalten als Voraussetzung für sicherheitsrelevante Analysen

Projektkurztitel: MDL

Projektbericht 01/09

Juli 2009

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Visuelle Verarbeitung

Zielsetzung

Ziel des Projektes MDL ist es aus multimedialen Inhalten Informationen zu extrahieren, um diese einem Data Mining System zuzuführen. Im Bereich der visuellen Verarbeitung wurden hierbei drei Ziele definiert: *Shot Boundary Detection, Face Detection/Recognition* und *Map Detection.* Unter Shot Boundery Detection versteht man den Process innerhalb von Videosequenzen Schnittgrenzen zu detektieren. Dies ist notwendig um eine zeitliche Segmentierung zu erhalten, die die Ausgangsbasis für weitere Verarbeitungsschritte darstellt. Das Ziel von Face Reconition/'Detection ist es im ersten Schritt Gesichter in Bildern zu finden (Detection) und diese dann in einem zweiten Schritt einer bekannten Person zuzuordnen (Recognition). Map Detection, das teilweise als wissenschaftliches Neuland angesehen werden kann, zielt darauf ab in Videobildern eingeblendeten Landkarten zu identifizieren, um so vorhandenen Inhalten auch geographische Information zuordnen zu können.

Anforderungen in MDL

Die Anforderungen an die in Betracht kommenden Algorithmen ergeben sich weitgehend aus der Natur der zu verarbeitenden Eingangsdaten. Es kann dabei von der Annahme ausgegangen werden, dass die zu verarbeitenden Eingangsdaten weitgehend Videosequenzen aus Nachrichtensendungen sind, die sich im Wesentlichen aus Moderationen, Berichten und Interviews zusammensetzen. Daher sind relativ statische Bedingungen zu erwarten, in denen sich im Speziellen Personen innerhalb einer Szene nicht allzu stark bewegen und meist frontal zu sehen sind – es sollen vorwiegend Gesichter von Sprechern und Interviewpartnern erkannt werden. Im Allgemeinen stellen sich für alle drei Zielapplikationen typische Probleme, die sich aus den zur Verfügung stehenden Daten ergeben. Da die Möglichkeit der Einflussnahme auf den Bildaufnahmeprozess fehlt, scheiden viele Ansätze aus, da z.B. durch die 2D Natur der Daten vielleicht bessere Methoden nicht verwendet werden können, die explizit 3D Information benötigen würden. Weiters müssen die verwendeten Methoden auch unter teilweise schwierigen Bedingungen, die sich aus Kamerarauschen, Kompressionsartefakte, niedriger Auflösung, etc. ergeben, funktionieren. Andererseits können diese Probleme durch einen Informationsgewinn, der sich aus der zeitlichen Dimension ergibt (z.B. Face Tracking), teilweise kompensiert werden.

Shot Boundary Detection

Einleitung

Als *Shot Boundary Detection* (SBD) bezeichnet man den Prozess digitale Filme und Videos in kleinere logische Einheiten, sogenannte *Shots,* zu segmentieren, um dadurch weitere visuelle Bearbeitungsschritte zu ermöglichen. Eine *Shot Boundary* oder *Transition* ist die Grenze bzw. der Übergang zwischen zwei *Shots*. Betrachtet man den ursprüngliche Entstehungsprozess eines Films oder Videos, dann ist es naheliegend einen *Shot* als Gruppierung von Einzelbildern anzusehen, die einer durchgängigen Kameraaufnahme entstammen. Unabhängig von der verwendeten *Shot Boundary* oder *Transition* zwischen einzelnen *Shots* werden diese vom menschlichen Betrachter als grundlegende logische Einheit eines Videos wahrgenommen [Smeaton et al, 2001].

Nach [Boreczky & Rowe, 1996, Smeaton et al., 2001, Osian & van Gool, 2004] können *Shot Transitions* grundsätzlich in zwei Gruppen eingeteilt werden: *Cuts* und *Gradual Transitions*. *Cuts* sind abrupte, direkte *Transitions,* die die Einzelbilder von *Shots* einfach aneinander reihen. *Gradual Transitions* sind weiche Übergänge, die Einzelbilder der einzelnen *Shots,* vermischen und können weiter unterschieden werden in *Dissolves* und *Fades*. *Dissolves* überblenden die Einzelbilder zwischen *Shots* während man unter *Fades Shot Boundaries* bezeichnet, die von schwarz einblenden bzw. auf schwarz ausblenden. Weitere *Gradual Transitions* wie *Wipes* oder andere Spezialeffekte sind von untergeordneter Bedeutung, da diese verhältnismäßig selten vorkommen [Smeaton et al., 2001]. Durch die unterschiedliche Ausprägung der verschiedenen *Transitions* werden einzelne auf den Typ spezialisierte Detektoren entworfen und diese dann fusioniert [Smeaton & Over, 2001-2007].

Kategorisierung

Nach der Systematik von [Yuan et al., 2007] lassen sich die bestehenden SBD Ansätze anhand von 3 Aspekten betrachten. Die Repräsentation des visuellen Inhalts, die Konstruktion eines Kontinuitätssignals und schließlich die Klassifikation dessen. Das Kontinuitätssignal drückt dabei die Ähnlichkeit zeitlich benachbarter Bilder aus, um den Rückschluss auf vorkommende *Shot Boundaries* zu ermöglichen. Für die Repräsentation des visuellen Inhalts gibt es verschiedenste Ansätze, die pixelbasiert [Kikukawa & Kawafuchi, 1992, Choubey & Raghavan, 1997, Zhang et al., 1995], mit Histogrammen [Gargi et al., 2000], Kanten [Zabih et al., 1995], mit Bewegung [Bouthemy et al., 1997] oder mit statistischen Maßzahlen wie Mittelwert und Standardabweichung von Intensitätswerten [Lienhart, 1999] arbeiten. Pixelbasierte Ansätze sind sensitiv gegen lokale oder globale Bewegung. Um die Invarianz zu erhöhen wurden verschiedene Varianten vorgeschlagen u.a. das Bild vorzuglätten. Mehr Invarianz gegen Bewegung bieten globale Farbhistogramme, die keine räumliche Information berücksichtigen. Dies kann aber bei *Shots,* die eine ähnliche Farbverteilung aufweisen, dazu führen, dass *Shot Boundaries* nicht erkannt werden. Eine Kombination von pixel-basierten und globalen Histogramm Methoden kann mit *Block-Matching* erreicht werden. Dabei wird jeder Frame in nicht überlappende Blöcke eingeteilt aus denen *Features* extrahiert und in weiterer Folge verglichen werden [Ahmed et al., 1999]. Komplexere Kanten-basierte Ansätze wie das Edge Change Ratio [Zabih et al, 1995] haben den Nachteil, dass sie deutlich rechenaufwändiger sind als einfache histogrammbasierte Ansätze. Weiters wurde gezeigt, dass diese sich nicht durchwegs besser für die SBD eignen [Lienhart, 1999]. Eine Ausnahme bildet die Überlegenheit bei Beleuchtungsänderungen wie etwa Blitzlichtern (*Flash-Lights*), dieser Nachteil wird aber von anderen Ansätzen mit speziellen *Flash* Detektoren [Kawai et al., 2008] kompensiert.

Basierend auf der Repräsentation des visuellen Inhalts wird das Kontinuitätssignal gebildet. Pixel-basierte Ansätze berücksichtigen die Differenz der Werte, Histogramm-basierte Ansätze greifen auf den χ^2 Test [Gargi et al, 2000] oder *Histogram Intersection* zurück. Kantenbasierte Methoden arbeiten u.a. mit Vergleichen von *Edge Maps* [Zabih et al., 1995]. Weiters kann für das Kontinuitätssignal z.B. über *Block-Matching* eine Invarianz bezüglich Kamera- bzw. Objektbewegung erreicht werden. Zwischen zwei sequentiellen Bildern wird dabei jedem

Pixelblock eine Entsprechung anhand eines Ähnlichkeitsmaßes im anderen Bild zugeordnet und die Kontinuität über das geeignetste Block-Paar basierend auf den jeweiligen *Features* bestimmt [Kawai et al., 2007]. Das gesamte Kontinuitätssignal kann im Wesentlichen nur den direkten Vergleich zwischen benachbarten Frames berücksichtigen oder auch Kontextinformation aus dem Vergleich mehrerer umliegender *Frames* miteinbeziehen. Die Klassifikation des Kontinuitätssignals erfolgt meist regelbasiert; es gibt jedoch Methoden, die auf Maschinellem Lernen beruhen. Bei einem regelbasierten Ansatz werden die Entscheidungsgrenzen, wann eine *Shot Boundary* auftritt, manuell modelliert. Die dabei verwendeten Schwellwerte können global gesetzt werden oder auch adaptiv sein [Volkmer et al., 2004, Truong et al., 2000, Yeo & Liu, 1995]. Hierbei wird der Schwellwert lokal in einem Detektionsfenster anhand des Kontinuitätssignals bestimmt. Zahlreiche Experimente haben jedoch gezeigt, dass adaptive Schwellwerte in der Praxis bessere Ergebnisse liefern [Lienhart, 2001, Hanjalic, 2002]. Andererseits versuchen SBD Ansätze, die auf Maschinelles Lernen zurückgreifen, die Entscheidungsgrenzen automatisch aus annotierten Daten zu lernen. Hierbei werden K-means [Naphade et al., 1998], KNN [Cooper, 2004] oder *Support Vector Machines* (SVMs) als zugrundeliegende Lernverfahren eingesetzt [Yuan et al., 2005, Ngo, 2003, Chua et al., 2003, Feng et al., 2005].

Übersicht über ausgewählte Methoden

Bedingt durch die große Anzahl an vorhandenen SBD Algorithmen und deren Bedeutung für die weitere Verarbeitungskette gibt es bereits frühzeitig Arbeiten, die sich mit der Evaluierung auf standardisierten Testdaten beschäftigen. Ein wesentlicher Meilenstein konnte dabei durch den ab 2001 jährlich stattfindenden TREC *Video Retrieval Evaluation* (TRECVID) Workshop gesetzt werden, der u.a. das Ziel verfolgt eine offene Evaluierungsplattform für SBD bereitzustellen. Dabei werden die Algorithmen der teilnehmenden Forschergruppen auf großen Testdatensätzen evaluiert, was einen seriösen und objektiven Vergleich ermöglicht [Smeaton et al. 2001, Smeaton et al., 2006]. Es ist daher naheliegend, die in Frage kommenden Algorithmen aufgrund ihrer TRECVID Ergebnisse zu vergleichen. Aufgrund der hohen Praxistauglichkeit der evaluierten Algorithmen wurde 2007 der *Shot Boundary Detection* Task erfolgreich geschlossen [Smeaton & Over, 2001-2007]. Wir widmen uns daher im Speziellen den (letzten) TRECVID Resultaten aus dem Jahre 2007. Der Testdatensatz enthält 18 Nachrichtenvideos mit 2463 Transitions (2236 Cuts, 134 Dissolves, 93 Anderen).

Der schnellste aller getesteten Ansätze, entworfen von NHK Science and Technical Research Laboratories [Kawai et al., 2008], verwendet separate, mehrstufig aufgebaute Detektoren für jeden Transitionstyp. Ziel ist es in den untergeordneten Stufen mit möglichst geringem Rechenaufwand potentielle Transitions-Kandidaten zu identifizieren und rechenaufwändigere Schritte nur auf diese anzuwenden. Dabei wird im ersten Schritt pixelbasiert gearbeitet; gegebenenfalls wird später auf *Block-Matching* zurückgegriffen. Die einzelnen Kontinuitätssignale der Detektoren variieren je nach Typ. Der Cut-Detektor arbeitet nur direkt mit den RGB Pixel bzw. Histogrammdifferenzen benachbarter Bilder, der Fade-Detektor analysiert den Luminanzverlauf der einzelnen Bilder im Kontext eines schwarzen Bildes, während der Dissolve-Detekor davon ausgeht, dass der Wert eines Pixels über einen beschränkten Zeitraum entweder monoton steigt oder sinkt. Die Klassifikation der einzelnen SB erfolgt schließlich regelbasiert über globale Schwellwerte. Die Laufzeit wird mit fünffacher Echtzeit angegeben.

Ein effizienter SVM-basierten Ansatz wurde in [Zhi-Cheng et al. 2008] vorgestellt. Der visuelle Inhalt wird über verschiedenste *Features* wie HSV/RGB Histogramme, Phasenhistogramm der Bewegungsvektoren, MFCC, und Luminanz repräsentiert. Das Kontinuitätssignal wird mit verschiedenen Distanzmetriken innerhalb eines *Sliding Windows* gebildet, wobei auch auf die in MPEG kodierte Information zurückgegriffen wird. Für die einzelnen Transitionstypen werden verschiedene Klassifikatoren mittels SVMs trainiert. Die SB Kandidaten werden dann mit zusätzlichen Informationen wie Kamera- und/oder Objektbewegung für die *Gradual Transition Detection* und dem *Edge Change* für die *Cut Detection* gefiltert. Die Laufzeit des Ansatzes liegt unter fünffacher Echtzeit.

Direkt in der *Compressed-Domain* des MPEG enkodierten Videos arbeitet der Ansatz in [Jinchang et al., 2008]. Hierbei werden die schon im MPEG Datenstrom enkodierte Informationen wie Bewegungsvektoren und andere lokale *Features* der Makroblöcke genutzt um SB Kandidaten zu identifizieren. Die Ähnlichkeit des Start- und Endbildes wird dann mittels Phasenkorrelation untersucht und regelbasiert entschieden. Auch hier bewegt sich die Laufzeit im Bereich von etwas unter fünffacher Echtzeit. Aufgrund der Modellierung ist der Ansatz natürlich auf MPEG Videos beschränkt.

Ein weiterer SVM basierter Ansatz ist [Liu et al., 2008]. Dabei werden pro Einzelbild *intra-Frame Features* wie RGB Histogramme und Kanten extrahiert. Davon werden diverse statistische *Features* abgeleitet. Weiters werden *inter-Frame Features* extrahiert die die bewegungskompensierten *Matching Errors* (*Block-Matching*) und Histogramm Änderungen betrachten, wobei die zeitliche Ableitung der *Features* geglättet wird. Um Randeffekte zu vermeiden, wird für die *Feature* Berechnung in den *Frames* jeweils nur eine zentrale Region of Interest (ROI) verwendet. Weiters hat der Ansatz einen eigenen Zoom-Detektor um *False Positives* zu filtern. Jeder der sechs verwendeten Detektortypen (jeweils für eine *Transition*) verwendet ein anderes Kontinuitätssignal basierend auf unterschiedlichen Featuretypen. Allein der Cut-Detektor greift auf 22 verschiedene Featuretypen zurück. Der Ansatz besitzt eine Laufzeit von 2-3 facher Echtzeit.

Der Ansatz von [Yuan et al., 2008] repräsentiert den visuellen Inhalt durch blockbasierte RGB Histogramme mit unterschiedlichen Blockgranularitäten für die individuellen Transition-Detektoren. Für den Cut-Detektor wird für die Klassifikation nur die Kontextinformation von zwei benachbarten Frames verwendet. Für die Detektion von *Gradual Transitions* wird mehr Kontextinformation einbezogen; diese wird auch zeitlich gesehen durch *Multi-Resolution* aufgelöst. Die Klassifikation erfolgt mittels zuvor trainierter SVMs. Um *False Positives* zu filtern werden potentielle *Gradual Transition* Kandidaten zusätzlich einer Kamera- und Objektbewegungsabschätzung mittels *Block-Matching* unterzogen. Abschließend wird sowohl für *Cut* und *Gradual Transition* Kandidaten mittels SIFT [Lowe, 1999] *Matching* überprüft, ob sich nicht gleiche Objekte vor bzw. nach der vermeintlichen SB befinden. Gegebenenfalls wird diese dann verworfen. Die Bewegungsabschätzung und das SIFT *Matching* arbeiten regelbasiert, was schließlich eine Echtzeit Laufzeit ermöglicht.

Zusammenfassung und Diskussion

Da die Laufzeit der Algorithmen einen wesentlichen Einfluss auf die Praxistauglichkeit hat, und um die Menge der für uns in Frage kommenden Algorithmen einzugrenzen, wurden nur Algorithmen betrachtet, die von der SBD Performance besser abschneiden als der schnellste evaluierte Algorithmus [Kawai et al. 2008]. Die SBD *Performance* wurde dabei mit dem F1 *Score* gemessen der sich aus dem harmonischen Mittel aus *Precision* und *Recall* bildet. Es ist anzumerken, dass der schnellste evaluierte Algorithmus einen höheren *F1 Score* für *Cuts* aufweist als alle Algorithmen der 2006 Evaluierung! Daher ist davon auszugehen, dass die für TRECVID 2007 untersuchten Algorithmen tendenziell bessere Ergebnisse liefern als ihre Vorgänger. Die Resultate der jeweiligen Algorithmen können Tabelle 1 entnommen werden.

Die Resultate der beschriebenen TRECVID Algorithmen zeigen, dass die komplexeren SVM-basierten Ansätze etwas besser abschneiden, aber durch den größeren Rechenaufwand längere Laufzeiten aufweisen. Betrachtet man die Gruppe der performanten Ansätze wie [Kawai et al. 2008], [Zhi-Cheng et al. 2008] und [Jinchang et al., 2008], so fällt auf, dass der F1 Score im Bereich der *Cuts* nur geringe Unterschiede aufweist hingegen [Kawai et al. 2008] bei den *Gradual Transitions* deutlich bessere Ergebnisse liefert. Im Vergleich zu den besseren SVM basierten Ansätzen von [Liu et al., 2008] und [Yuan et al., 2008] erreicht [Kawai et al. 2008] einen um 8-10% schlechteren F1 Score für *Gradual Transitions*. Bezüglich der Laufzeiten ergibt sich aber ein Geschwindigkeitsvorteil von Faktor 2—5! Weiters müssen die rein regelbasierten Ansätze nicht trainiert werden.

Es sei darauf hingewiesen, dass SBD Publikationen, die nicht mit TRECVID in Verbindung stehen wie [Osian & van Gool 2004, Ren & Singh 2004, Feng et al. 2005] ebenfalls vergleichbare Resultate liefern. Durch unterschiedliche Testdaten fehlt dadurch aber eine objektive Vergleichsbasis, und eine Einordnung der Resultate ist schwierig. Die verwendeten Methoden ähneln im Wesentlichen denen der beschriebenen TRECVID Algorithmen.

	Summe			Cut			Gradual Transition		
Teilnehmer	R	P	F1	R	P	F1	R	P	F1
[Kawai et al. 2008]	0.91	0.94	0.93	0.93	0.97	0.95	0.61	0.69	0.65
[Zhi-Cheng et al. 2008]	0.90	0.93	0.92	0.96	0.97	0.97	0.24	0.36	0.30
[Jinchang et al., 2008]	0.94	0.92	0.93	0.97	0.98	0.98	0.59	0.43	0.49
[Liu et al., 2008]	0.96	0.95	0.96	0.98	0.97	0.97	0.71	0.80	0.75
[Yuan et al., 2008]	0.95	0.96	0.95	0.97	0.98	0.98	0.72	0.73	0.73

Tabelle 1: TRECVID 2007 ausgewählte Resultate aufsteigend sortiert nach der Laufzeit.

Literaturverzeichnis

[Ahmed et al., 1999] Ahmed, M., Karmouch, A., und Abu-Hakima, S. (1999). Key frame extraction and indexing for multimedia databases. In *Proc. Vision Interface Conf.*, Seiten 506–511.

[Boreczky & Rowe, 1996] Boreczky, J. und Rowe, L. (1996). Comparison of video shot boundary detection techniques. In *Storage and Retrieval for Still Image and Video Databases IV*.

[Bouthemy et al., 1997] Bouthemy, P., Gelgon, M., und Ganansia, F. (1997). A unified approach to shot change detection and camera motion characterization. *IEEE Trans. on Circuit and Systems for Video Technology*, 9:1030–1044.

[Choubey & Raghavan, 1997] Choubey, S. K. und Raghavan, V. V. (1997). Generic and fully automatic content based image retrieval architecture. *Pattern Recognition Letters*, 18:11–13.

[Chua et al., 2003] Chua, T., Feng, H., und Chandrashekhara, A. (2003). An unified framework for shot boundary detection via active learning. In *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, Band 2, Seite 845.

[Cooper, 2004] Cooper, M. (2004). Video segmentation combining similarity analysis and classification. In *Proc. ACM Intern. Conf. on Multimedia*, Seiten 252–255.

[Feng et al., 2005] Feng, H., Fang, W., Liu, S., und Fang, Y. (2005). A new general framework for shot boundary detection and key-frame extraction. In *Proc. ACM Intern. Workshop on Multimedia Information Retrieval*, Seiten 121–126.

[Gargi et al., 2000] Gargi, U., Kasturi, R., und Strayer, S. (2000). Performance characterization of video-shot-change detection methods. *IEEE Trans. on Circuit and Systems for Video Technology*, 10(1):1–13.

[Hanjalic, 2002] Hanjalic, A. (2002). Shot-boundary detection: unraveled and resolved? *IEEE Trans. on Circuit and Systems for Video Technology*, 12(2):90–105.

[Jinchang et al., 2008] Jinchang, R., Jianmin, J., und Juan, C. (2008). Determination of shot boundary in MPEG videos for TRECVID 2007. In *Proc. TREC Video Retrieval Evaluation Workshop*. National Institute of Standards and Technology.

[Kawai et al., 2008] Kawai, Y., Sumiyoshi, H., und Yagi, N. (2008). Shot boundary detection at TRECVID 2007. In *Proc. TREC Video Retrieval Evaluation Workshop*. National Institute of Standards and Technology.

[Kikukawa & Kawafuchi, 1992] Kikukawa, T. und Kawafuchi, S. (1992). Development of an automatic summary editing system for the audio visual resources. Technical Report, Institute of Electronics, Information and Communication Engineers.

[Lienhart, 1999] Lienhart, R. (1999). Comparison of automatic shot boundary detection algorithms. In *Proc. Storage and Retrieval for Image and Video Databases*, Seiten 290–301.

[Lienhart, 2001] Lienhart, R. (2001). Reliable transition detection in videos: A survey and practitioners guide. *Intern. Journal of Image and Graphics*, 1:469–486.

[Liu et al., 2008] Liu, Z., Zavesky, E., Gibbon, D., Shahraray, B., und Haffner, P. (2008). AT&T research at TRECVID 2007. In *Proc. TREC Video Retrieval Evaluation Workshop*. National Institute of Standards and Technology.

[Naphade et al., 1998] Naphade, M., Mehrotra, R., Ferman, A., Warnick, J., Huang, T., und Tekalp, A. (1998). A high-performance shot boundary detection algorithm using multiple cues. In *Proc. IEEE Intern. Conf. on Image Processing*, Band 1, Seiten 884–887.

[Ngo, 2003] Ngo, C. (2003). A robust dissolve detector by support vector machine. In *Proc. ACM Intern. Conf. on Multimedia*, Seiten 283–286. ACM.

[Osian & van Gool, 2004] Osian, M. und van Gool, L. (2004). Video shot characterization. *Machine Vision and Applications*, 15(3):172–177.

[Ren & Singh, 2004] Ren, W. und Singh, S. (2004). Automatic video shot boundary detection using machine learning. In *Proc. Intern. Conf. on Intelligent Data Engineering and Automated Learning*, Seiten 285–292. Springer.

[Smeaton & Over, 2007] Smeaton, A. F. und Over, P. (2001-2007). TRECVID: shot boundary detection task summaries.

http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[Smeaton et al., 2006] Smeaton, A. F., Over, P., und Kraaij, W. (2006). Evaluation campaigns and TRECVID. In *Proc. ACM Intern. Workshop on Multimedia Information Retrieval*, Seiten 321–330.

[Smeaton et al., 2001] Smeaton, A. F., Taban, R., und Over, P. (2001). The TREC-2001 video track report. In *Proc. Text REtrieval Conf.*, Seite 52.

[Truong et al., 2000] Truong, B. T., Dorai, C., und Venkatesh, S. (2000). New enhancements to cut, fade, and dissolve detection processes in video segmentation. In *Proc. ACM Intern. Conf. on Multimedia*, Seiten 219–227.

[Volkmer et al., 2004] Volkmer, S., Tahanghoghi, M., und Williams, H. (2004). RMIT University at TRECVID 2004. In *Proc. TREC Video Retrieval Evaluation Workshop*.

[Yeo & Liu, 1995] Yeo, B. und Liu, B. (1995). Rapid scene analysis on compressed video. *IEEE Trans. on Circuit and Systems for Video Technology*, 5(6):533–544.

[Yuan et al., 2008] Yuan, J. et al. (2008). THU and ICRC at TRECVID 2007. In *Proc. TREC Video Retrieval Evaluation Workshop*. National Institute of Standards and Technology.

[Yuan et al., 2005] Yuan, J., Li, J., Lin, F., und Zhang, B. (2005). A unified shot boundary detection framework based on graph partition model. In *Proc. ACM Intern. Conf. on Multimedia*, Seiten 539–542.

[Yuan et al., 2007] Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., und Zhang, B. (2007). A formal study of shot boundary detection. *IEEE Trans. on Circuit and Systems for Video Technology*, 17:168–186.

[Zabih et al., 1995] Zabih, R., Miller, J., und Mai, K. (1995). A feature-based algorithm for detecting and classifying scene breaks. In *Proc. ACM Intern. Conf. on Multimedia*, Seiten 189–200.

[Zhang et al., 1995] Zhang, H., Low, C. Y., und Smoliar, S. W. (1995). Video parsing and browsing using compressed data. *Multimedia Tools and Applications*, 1(1):89–111.

[Zhi-Cheng et al., 2008] Zhi-Cheng, Z., Xing, Z., Tao, L., und An-Ni, C. (2008). BUPT at TRECVID 2007: shot boundary detection. In *Proc. TREC Video Retrieval Evaluation Workshop*. National Institute of Standards and Technology.

Face Detection

Einleitung

Das Ziel der *Face Detection* ist es, in beliebigen Bildern die Präsenz einer vorab unbekannten Anzahl von Gesichtern zu bestimmen. Außerdem soll die Position und Größe für jedes Gesicht ermittelt werden. *Face Detection* bildet dabei die Ausgangsbasis für weitere Verarbeitungsschritte wie etwa *Facial Feature Detection*, die Gesichtsteile detektiert oder *Face Recognition,* deren Ziel es ist spezifische Personen zu identifizieren. Die wesentlich Herausforderung in der *Face Detection* liegt in der natürlichen Variabilität der Pose (Frontal, Profil), der Orientierung, der Größe und der Position von Gesichtern, sowie weiteren Faktoren wie Gesichtsausdruck, Verdeckung und Beleuchtungseigenschaften, die sich teilweise sehr drastisch auf das Erscheinungsbild von Gesichtern auswirken [Yang et al., 2002].

Kategorisierung

Die vielfältigen Ansätze für *Face Detection* können nach [Yang et al., 2002] in auf menschlichem Vorabwissen basierende, *Feature*-invariante, auf *Template Matching*-basierende und *Appearanc-based* Methoden unterschieden werden. Methoden wie [Yang & Huang, 1994] arbeiten *top-down* regelbasiert mit enkodiertem menschlichen Vorabwissen, das es erlaubt Gesichter zu charakterisieren. Dabei wird der Zusammenhang zwischen den einzelnen Gesichtsteilen (*Facial Features)* modelliert. *Feature*-invariante Methoden [Dai & Nakano, 1996, Kjeldsen & Kender, 1996, Leung et al., 1995, McKenna et al., 1998] versuchen *bottom-up* diese *Facial Features* wie Augen(brauen), Nase, Mund und andere *Features* wie Hautfarbe etc. zu detektieren und dann mittels statistischem Modell die Präsenz eines Gesichtes abzuleiten und zu verifizieren. Obwohl die Mehrzahl der wissensbasierten und *Feature*-invarianten Methoden sich auf den Terminus *Face Detection* beziehen, ist in den meisten Fällen nur die Lokalisierung eines einzelnen Gesichts gemeint, bei dem bereits bekannt ist, dass es sich im Bild befindet. Auf *Template Matching* besierende Methoden arbeiten mit manuel vordefinierten und vorkonfigurierten generischen Gesichtsschablonen. Die Präsenz eines Gesichts wird über die Korrelation der im Suchbild vorkommenden Muster mit den im *Template* gespeicherten verifiziert. Dabei werden verschiedenste Eigenschaften wie etwa die Gesichtskontur oder die einzelnen *Facial Features* unabhängig voneinander betrachtet. Um die Invarianz, gegenüber den vielfältigen Erscheinungsformen eines Gesichtes zu erhöhen, wurden verschiedenste Erweiterungen wie *Multiresolution* [Miao et al., 1999], *Subtemplates* [Sakai et al., 1969, Craw et al., 1992] und *elastische Templates* [Yuille et al., 1989] vorgeschlagen. Basierend auf statistischen Analysen und Maschinellem Lernen versuchen *Appearance-based* Methoden Modelle aus Trainingsdaten, die die natürliche Variabilität in den Daten abdecken, zu lernen. Dabei werden die Eigenschaften und Charakteristika, die positive wie negative Beispiele beschreiben, entweder in Form eines Verteilungsmodels gelernt (generatives Modell) oder durch eine diskkriminative Entscheidungsfunktion beschrieben (diskkriminatives Model). Im Allgemeinen, liefern *Appearance-based* Methoden im Vergleich mit den anderen genanten Ansätzen deutlich bessere Erkennungs- und Fehldetektionsraten.

Übersicht über ausgewählte Methoden

Der erste richtungsweisenden *Appearance-based* Ansatz wird in [Rowley et al., 1996] beschrieben. Dabei werden verschiedene Neuronale Netze für frontale Gesichter unter der Verwendung von positiven und negativen Beispielen trainiert. Für die Detektion von Gesichtern in Testbildern basiert auf *Exhaustive Search,* wobei mit einem *Sliding Window* (Detektionsfenter) die gelernten Filter an jeder Position auf mehreren Skalierungen (die Größe des Bildes wird so angepasst, dass der gelernte Filter angewendet werden kann) angewendet werden. Die Neuronalen Netze setzen sich aus rezeptiven Feldern, die aus Block- und Streifenanordnungen mit verschiedenen Granularitäten aufgebaut sind, zusammen. Insgesamt wurden für das Training 1.050 positive Beispiele verwendet. Die notwendigen negativen Beispiele werden in der Trainingsphase mittels *Bootstrapping* unter der Verwendung von (gesichterten) Fehldetektionen gesammelt um einerseits auch diese

schwierigen Beispiele mithinzubeziehen und andererseits die manuelle Auswahl von geeigneten Gegenbeispielen zu vereinfachen. Die Autoren berichten von einer Detektionsrate von 84,4% bei nur 16 Fehldetektionen auf dem CMU-MIT Datensatz (siehe Appendix). Auf dem FERET Datensatz, (siehe Appendix), der eine geringere Komplexität aufweist und im Wesentlichen aus Portraits besteht, beträgt die Detektionsrate 99,2% bei 8 Fehldetektionen. Auf damaliger Hardware betrug die Auswertungszeit für ein Bild der Größe 320x240 Pixel etwas mehr als 6 Minuten.

Der Ansatz von [Moghaddam & Pentland, 1997] verfolgt eine modellbasierte probabilistische Herangehensweise für das Problem der frontalen *Face Detection* durch eine Analyse des Eigenspektrums und kann daher als Generalisierung der *Eigenfaces* Methode [Kirby & Sirovich 1990; Turk & Pentland 1991] betrachtet werden. Details zu Eigenfaces können aus Abschnitt 3 entnommen werden. Um ein Modell für die Wahrscheinlichkeitsdichte der einzelnen Klassen (Gesichter, *Facial Features*) zu erhalten wird entweder eine multivariate Gaußverteilung oder ein Gaußsches *Mixture Model* (GMM) herangezogen. Die Klassifikation erfolgt mittels *maximum-likelihood* Abschätzung, die sowohl die Distanz im *Feature Space* als auch den residualen Fehler berücksichtigt. Um die Robustheit zu erhöhen werden mit dem selben Ansatz zusätzlich innerhalb des Suchfensters *Facial Features* (Augen, Nase, Mund) detektiert. Die Autoren berichten von einer Detektionsrate von 97% auf dem FERET Datensatz. Zur Laufzeit des Algorithmus wird keine Angabe gemacht.

Schneidermann und Kanade widmen sich in [Schneiderman & Kanade, 2000] erstmals dem Problem der *Multi View Face Detection* (MFVD), der Detektion von Gesichtern auch in nicht frontalen Posen bis hin zum Profil. Allerdings wird davon ausgegangen, dass die Gesichter nicht geneigt sind. Anhand der spezifischen Gesichtspose (Frontal und rechtes Profil) werden zwei separate Detektoren trainiert die auf einem Detektionsfenster arbeiten. Der Detektor für das linke Profil wird über spiegeln des rechten Detektors erzeugt. Innerhalb des Detektionsfensters wird der visuelle Inhalt über quantifizierte *Wavelet* Koeffizienten in Form von Histogrammen repräsentiert. Ein Histogramm, das eine kleine Menge visueller Information darstellt, ist dabei eine multivariate Verteilung zwischen dem visuellen Attribut und der Position unter der Bedingung der positiven oder negativen Gesichtsklasse. Die Abschätzung der Wahrscheinlichkeitsdichte erfolgt über simples Zählen wie oft ein gewisses Attribut an einer bestimmen Position in den Trainingsdaten für die jeweilige Klasse vorkommt. Der Detektor setzt sich dann aus den jeweiligen Verteilungen zusammen, die die Variationen in der visuellen Erscheinung abdecken sollen. In der Trainingsphase werden für jedes der 2.000 Gesichtsbeispiele 400 synthetische Variationen erzeugt. Weiters wird auf *Bootstrapping* und *Adaboost* zurückgegriffen, um den vorläufigen Detektor inkrementell zu trainieren. Die Autoren berichten von einer Detektionsrate von 90.2% bei 110 Fehldetektionen für frontale Gesichter und 86.4% Detektionsrate bei 91 Fehldetektionen für Profile. Die Auswertung eines 320x240 Bildes nimmt, auf damaliger Hardware, in etwa 1 Minute in Anspruch.

Der Durchbruch in der frontalen *Face Detection* in Echtzeit gelang mit [Viola & Jones, 2001]. Den Grundstein für die Effizienz des Ansatzes liefert die aufgegriffene Bildrepräsentation, das *Integral Image,* die es erlaubt die verwendeten Rechtecksfeatures (Haar-Wavelets) mit sehr wenigen Operationen zu berechnen. Da das vorgeschlagene Detektionsfenster von 24x24 Pixel über 180.000 dieser *Features* enthält ist an eine Auswertung aller, unter Berücksichtigung von Effizienzgründen, nicht zu denken. Um ein schlankes *Feature Set* aus den Trainingsdaten zu lernen wird eine leicht modifizierte Version des *AdaBoost* Algorithmus verwendet (*Boosting for Feature Selection*). Die binäre Entscheidungsfunktion eines *Weak Classifiers* (WC) berücksichtigt dabei nur ein einzelnes *Feature*. Unter einem *Weak Classifier* versteht man einen Klassifikator, der im Falle von zwei Klassen nur eine (gering) bessere Klassfikationsrate als 50% aufweisen muss. Die eigentliche Auswahl der *Features* wird schließlich durch *Boosting* bestimmt. Um den Rechenaufwand gering zu halten wird in der vorgeschlagenen sequentiellen Filterkaskade möglichst früh versucht unwahrscheinliche Kandidaten zurückzuweisen. Dabei werden in den ersten Ebenen geboostete Klassifikatoren verwendet, die sich aus wenigen WC zusammensetzen und daher sehr effizient zu berechnen sind. Die überwiegende Mehrheit der Suchfenster wird in der ersten oder zweiten Ebene zurückgewiesen. Mit zunehmender Höhe in der Filterkaskade werden die Klassifikatoren komplexer. Der Ansatz erreicht auf dem CMU-MIT Testdatensatz eine Detektionsrate von 81,1% mit nur 10 Fehldetektionen bzw. 89% mit 31 Fehldetektionen. Auf einem 384x288 Bild erreicht der Detektor 15 Bilder pro Sekunde und damit die Ergebnisse von [Rowley et al., 1996, Schneiderman & Kanade, 2000] mit deutlich kürzerer Rechenzeit. Viele weitere nennenswerte Ansätze basieren auf Erweiterungen und Anleihen

von [Viola & Jones, 2001]; auch wird dieser Ansatz in vielen handelsübliche Digitalkameras eingesetzt [Ren et al., 2008].

Im Gegensatz zur Evaluierung eines Detektionsfensters wird in [Schneiderman, 2004] eine *Feature* orientierte Auswertungsmethode vorgeschlagen. Das Ziel ist es die erste Ebene der Filterkaskade robuster zu gestalten, da davon ausgegangen werden kann, dass einzelne Features nicht verlässlich genug sind, um alle positiven Beispiele zu klassifizieren und gleichzeitig den überwiegenden Teil der negativen Beispiele zurückzuweisen. Dabei werden die regulär gesampelten Features und deren Auswertungen zwischen den Klassfikationsfunktionen der überlappenden Detektionsfenster geteilt. Da jedes Feature nur einmal ausgewertet wird, ist nur die komplexere Klassifikationsfunktion als zusätzlicher Aufwand zu sehen. Der Ansatz vereint im wesentlichen jenen in [Schneiderman & Kanade, 2000] mit der Detektorstruktur in [Viola & Jones, 2001]. Die Autoren berichten von besseren Detektionsraten als [Viola & Jones, 2001] auf frontalen Gesichtsbeispielen und einer Detektionsrate von 87% bei 86 Fehldetektionen auf dem CMU Test Set für nicht frontale *Face Detection*. Dies ist leicht besser als [Schneiderman & Kanade, 2000]. Der Ansatz erreicht auf damaliger Hardware bei einer Auflösung von 300x200 5 Bilder die Sekunde bei frontaler Gesichtsdetektion und 1 Bild pro Sekunde bei MFVD.

Mittels *FloatBoost* kombiniert [Zhang, 2004] *Floating Search* und *AdaBoost*. *FloatBoost* verwendet einen *backtrack* Mechanismus nach jeder *AdaBoost* Iteration, um WCs zu entfernen, die sich nicht auf die Fehldetektionsrate auswirken. Weiters wird eine pyramidenartige Klassifikationsstruktur für MVFD vorgestellt. Dabei wird die Ansichtsmenge der Gesichtsposen auf jeder Pyramidenebene in immer feinere Untergruppen geteilt, die dann von speziellen Klassifikatoren behandelt werden. Wird ein Detektionskandidat von einem speziellen Klassifikator in der Pyramide zurückgewiesen, wird dieser an alle weiteren in der jeweiligen Ebene weitergereicht. Die einzelnen Klassifikatoren sind selbst wieder eine sequentielle Filterkaskade wie in [Viola & Jones, 2001]. Die Autoren berichten von einer niedrigeren Fehldetektionsrate im Vergleich zu *AdaBoost*, oder weniger WCs bei gleicher Fehldetektionsrate. Die MVFD arbeitet auf damaliger Hardware mit 5 Bildern pro Sekunde, allerdings werden keine Resultate für die MVFD angegeben.

Um die Diskriminativität der *Features* zu erhöhen wurde in [Wu et al., 2004] unter Verwendung von *Real Adaboost* [Schapire et al., 1999] der Ansatz von Viola&Jones mit einem *confidence-rated look-up-table* (LUT) für Haar *Features* erweitert. *Real Adaboost* basiert auf WC deren Auswertungsfunktion reelle Werte liefert, im Kontrast zu diskreten. Verglichen mit *Adaboost* soll dies zu einer schnelleren Konvergenz und zu einer diskriminativeren Beschreibung der Daten führen. Der vorgeschlagene LUT teilt daher die Auswertungsfunktion der Haar *Features* in mehrere Abschnitte ein, um jedem dieser eine Konfidenz für die Präsenz und Absenz des Objektes zuzuordnen. Weiters wurde die Filterkaskade modifiziert, um die Konfidenz der Klassifikation einer Ebene in die nächste Klassifikationsfunktion zu propagieren. Im Wesentlichen wird dies durch die Aufnahme des letzten Klassifikators als WC in die neue *Boosting* Iteration realisiert. Für MVFD setzt der Ansatz auf spezialisierte parallele Detektoren, die jeweils durch eine eigene Filterkaskade repräsentiert sind. Zur Laufzeitoptimierung wird vorgeschlagen in den ersten Ebenen parallel eine *Pose Estimation* durchzuführen und dann nur die vielversprechendste Hypothese weiter auszuwerten. Der Ansatz bietet auf dem CMU-MIT Testdatensatz für frontale Gesichter eine Detektionsrate von 90.1% bei 10 Fehldetektionen. Für die MVFD auf dem CMU Profil Testdatensatz erreicht der Ansatz 84.8% bei 34 Fehldetektionen. Die Laufzeit für MVFD liegt, auf damaliger Hardware, bei etwa 12 Bildern pro Sekunde bei einer Auflösung von 320×240.

Der Ansatz von [Huang et al., 2005] zielt auf eine fundamentale Verbesserung der Struktur der bisherigen Filterkaskaden ab. Der *Width-First-Search* (WFS) Baum versucht gleichzeitig die Unterschiede und Gemeinsamkeiten zwischen den einzelnen Gesichtsposen zu modellieren. *Vector Boost*, eine Erweiterung von *Real Adaboost*, zerlegt dabei das komplexe (multi-class) Problem, die einzelnen Posen vom Hintergrund bzw. auch untereinander zu unterschieden, in einzelne binäre Entscheidungsprobleme. Der wesentliche Unterschied liegt darin, dass individuell selektierte *Features* innerhalb des *Vector Boosting Frameworks* zwischen den einzelnen Klassifikatoren geteilt und so die Gemeinsamkeiten besser berücksichtigt werden können als bei individuellen binären Klassifikationsfunktion. Weiters wird nicht zwingend exklusiv klassifiziert, was sich vorteilhaft für komplexe Beispiele erweist. Im WFS Baum wird in jedem Baumknoten ein *Vector Boost* Klassifikator eingebettet. Dieser deckt dann einen gewissen Bereich über die Menge der Gesichtsposen ab. Dabei werden die wahrscheinlichen Kandidaten an die jeweiligen geeigneten weiteren Baumknoten propagiert

um eine finale Entscheidung zu erhalten. Der fertig aufgebaute WFS MVFD Baum in [Huang et al., 2005] setzt sich aus 204 Knoten in 16 Ebenen und 7081 WCs zusammen. Dabei wurden 75.000 Gesichtsbeispiele trainiert. Die Autoren berichten Ergebnisse vom CMU Profil Testdatensatz wobei der Ansatz mit einer Detektionsrate von 83% bei 16 Fehldetektionen aufwartet.

Einen auf *Error Correcting Output Codes* (ECOC) [Dietterich & Bakiri, 1995] basierten MVFD Ansatz wird in [Zhang et al., 2006] verfolgt. Das MVFD Klassifikationsproblem wird in Form von mehreren binären Basisklassifikatoren modelliert, die in Summe einen gemeinsamen Klassifikator bilden, der sich aus binären ECOCs zusammensetzt. Aufgrund dessen können Falschklassfikationen erkannt und korrigiert werden. Für die Basisklassifikatoren werden dabei räumliche Histogramm *Features* in einer Filterkaskade verwendet. Diese werden mit dem Fisher Kriterium anhand ihrer Diskriminierbarkeit in der Trainingsphase iterativ ausgewählt. Nach der groben Filterkaskade folgt eine SVM-basierte Klassifikation. Die Autoren berichten von einer Detektionsrate von 82% bei 91 Fehldetektionen auf dem CMU Profil Testdatensatz.

Mittels S*parse Features* im *Granular Space* und einer heuristischen Suche für die *Feature* Selektion erweitert [Huang et al., 2007] im Wesentlichen [Huang et al., 2005]. Dabei wird die Idee verfolgt komplexere irreguläre Muster besser abzubilden als die originalen Haar *Features*. Gleichzeitig soll der Rechenaufwand in etwa gleich bleiben. Der *Granular Space* besteht aus der Menge der granularen Bildern die durch einen Glättungsfilter mit ansteigendem *Scale* erzeugt werden. Dabei ist eine *Granule* definiert durch den x,y *Offset* und den *Scale* des Filters. Die *Features* setzen sich aus einer Linearkombination von (nur sehr wenigen) *Granules* zusammen. Durch die große Anzahl möglicher Kombinationen für die *Sparse Features* wird auch eine spezielle Selektionsstrategie, die heuristische Suche, vorgeschlagen. Die Autoren berichten von besseren Ergebnissen als [Huang et al., 2005] auf dem CMU Profil Testdatensatz.

Zusammenfassung und Diskussion

Zusammenfassend kann gesagt werden, dass existierende *Face Detection* Ansätze durchwegs gut für die frontale *Face Detection* geeignet sind. Abbildung 1 verdeutlicht dies durch die Resultate der einzelnen Methoden auf dem CMU-MIT Testdatensatz, bestehend aus 130 Bildern, die 507 frontale Gesichter enthalten. Mit [Viola & Jones, 2001] wurde ein Ansatz vorgestellt der erstmals Echtzeitfähigkeit bei ansprechender Erkennungsrate zeigt und gleichzeitig den Auslöser für die intensive Forschung an MVFD Ansätzen bildet. Diese erreichen vielversprechende Resultate, was auch aus Abbildung 2 hervorgeht. Die Methoden wurden auf dem CMU Datensatz für nicht-frontale Gesichtsdetektion evaluiert. Der Datensatz enthält 441 Gesichter in 208 Bildern, wovon 347 als Profilaufnahmen annotiert sind.

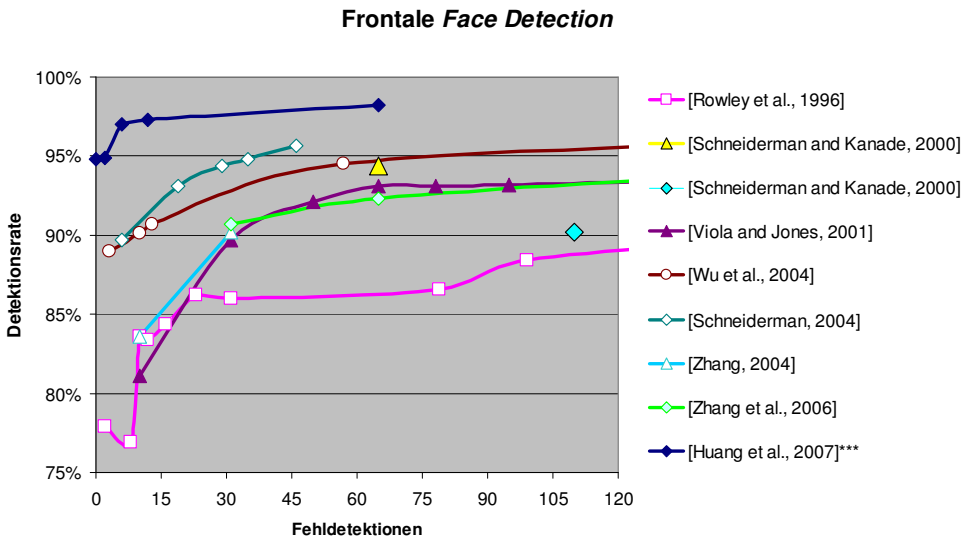

Abbildung 1 Frontale *Face Detection* Ansätze CMU-MIT Testdatensatz. ***Die berichteten Resultate wurden aus einem Diagramm aus [Huang et al., 2007] entnommen.

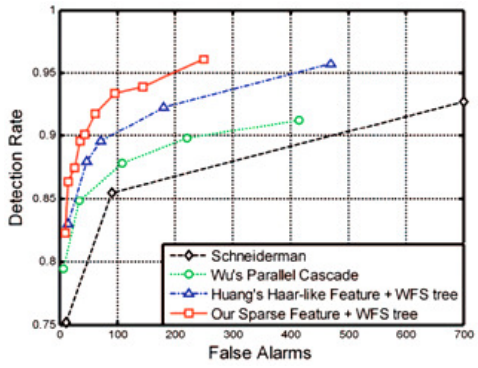

Abbildung 2 MVFD Ansätze CMU Profil Testdatensatz [Huang et al., 2007].

Literaturverzeichnis

[Craw et al., 1992] Craw, I., Tock, D., und Bennett, A. (1992). Finding face features. In *Proc. European Conf. on Computer Vision*, Band 3954, Seiten 92–96.

[Dai & Nakano, 1996] Dai, Y. und Nakano, Y. (1996). Face-texture model-based on SGLD and its application in face detection in a color scene. *Pattern Recognition*, 29(6):1007–1017.

[Dietterich & Bakiri, 1994] Dietterich, T. G. und Bakiri, G. (1994). Solving multiclass learning problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research*, 2:263–286.

[Huang et al., 2005] Huang, C., Ai, H., Li, Y., und Lao, S. (2005). Vector boosting for rotation invariant Multi-View face detection. In *Proc. IEEE Intern. Conf. on Computer Vision*, Seiten 446–453.

[Huang et al., 2007] Huang, C., Ai, H., Li, Y., und Lao, S. (2007). High-Performance rotation invariant multiview face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):671–686.

[Kirby & Sirovich, 1990] Kirby, M. und Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1):103–108.

[Kjeldsen & Kender, 1996] Kjeldsen, R. und Kender, J. (1996). Finding skin in color images. In *Proc. IEEE Intern. Conf. on Automatic Face and Gesture Recognition*, Seite 312–317.

[Leung et al., 1995] Leung, T., Burl, M., und Perona, P. (1995). Finding faces in cluttered scenes using random labeled graph matching. In *Proc. IEEE Intern. Conf. on Computer Vision*, Seiten 637–644.

[McKenna et al., 1998] McKenna, S., Gong, S., und Raja, Y. (1998). Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892.

[Miao et al., 1999] Miao, J., Yin, B., Wang, K., Shen, L., und Chen, X. (1999). A hierarchical multiscale and multiangle system for human face detection in a complex background using gravity-center template. *Pattern Recognition*, 32(7):1237–1248.

[Moghaddam & Pentland, 1997] Moghaddam, B. und Pentland, A. (1997). Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):696–710.

[Nechyba et al., 2008] Nechyba, M., Brandy, L., und Schneiderman, H. (2008). PittPatt face detection and tracking for the CLEAR 2007 evaluation. In *Multimodal Technologies for Perception of Humans*, Seiten 126–137.

[Turk & Pentland, 1991] Turk, M. und Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

[Ren et al., 2008] Ren, J., Kehtarnavaz, N., und Estevez, L. (2008). Real-time optimization of Viola-Jones face detection for mobile platforms. In *IEEE Workshop on Circuits and Systems: System-on-Chip - Design, Applications, Integration, and Software*, Seiten 1–4.

[Rowley et al., 1996] Rowley, H., Baluja, S., und Kanade, T. (1996). Neural network-based face detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 203–208.

[Sakai et al., 1969] Sakai, T., Nagao, M., und Fujibayashi, S. (1969). Line extraction and pattern detection in a photograph. *Pattern Recognition*, 1(3):233–236.

[Schapire & Singer, 1999] Schapire, R. E. und Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336.

[Schneiderman, 2004] Schneiderman, H. (2004). Feature-centric evaluation for efficient cascaded object detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Band 2, Seiten 29–36.

[Schneiderman & Kanade, 2000] Schneiderman, H. und Kanade, T. (2000). A statistical method for 3D object detection applied to faces and cars. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Band 1, Seiten 746–751.

[Viola & Jones, 2001] Viola, P. und Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Band 1, Seiten 511–518.

[Viola & Jones, 2004] Viola, P. und Jones, M. J. (2004). Robust Real-Time face detection. *Intern. Journal of Computer Vision*, 57(2):137–154.

[Wu et al., 2004] Wu, B., Ai, H., Huang, C., und Lao, S. (2004). Fast rotation invariant multi-view face detection based on real adaboost. In *Proc. IEEE Intern. Conf. on Automatic Face and Gesture Recognition*, Seiten 79–84.

[Yang & Huang, 1994] Yang, G. und Huang, T. S. (1994). Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63.

[Yang et al., 2002] Yang, M., Kriegman, D., und Ahuja, N. (2002). Detecting faces in images: a survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58.

[Yuille et al., 1989] Yuille, A., Cohen, D., und Hallinan, P. (1989). Feature extraction from faces using deformable templates. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 104–109.

[Zhang et al., 2006] Zhang, H., Gao, W., Chen, X., Shan, S., und Zhao, D. (2006). Robust multi-view face detection using error correcting output codes. In *Proc. European Conf. on Computer Vision*, Band 3954/2006, Seiten 1–12.

[Zhang, 2004] Zhang, Z. (2004). FloatBoost learning and statistical face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9).

Face Recognition

Definition / Abgrenzung / Anforderungen, Schwierigkeiten

Als *Face Recognition* bezeichnet man die Aufgabe in digitalen Bildern (Einzelbilder/Fotos oder Bildsequenzen/Videos) dem System bekannte menschliche Gesichter zu finden und zu erkennen. Das menschliche Wahrnehmungssystem hat gerade in diesem Bereich außergewöhnliche Fähigkeiten – Untersuchungen legen nahe, dass es im Gehirn eigens auf diese Aufgabe spezialisierte Areale gibt [Biederman & Kalocsai 1998; Ellis1986; Gauthier et al. 1999; Gauthier & Logothetis 2000] und trotz intensiver Forschung und großer Erfolge gerade in den letzten Jahren scheint es im Moment noch utopisch zu sein, die Gesamtheit dieser Leistungen mit computergestützten automatischen System jemals imitieren zu können. Dennoch wurden in den letzten zwei Jahrzehnten große Fortschritte in einigen Teilbereichen erzielt, sodass heute bereits kommerzielle System angeboten werden können, die jeweils eine definierte Aufgabe zufriedenstellend lösen.

Im Allgemeinen durchläuft ein *Face Recognition* Prozess drei Stufen. Den ersten Schritt bildet dabei die *Face Detection*, wie im vorangegangenen Abschnitt beschrieben. Anschließend verfolgt *Facial Feature Detection* das Ziel den genauen Ort gewisser Charakteristika des Gesichts (Augen, Nase, Mund...) zu bestimmen und durch geeignete mathematische Modelle zu beschreiben. In der letzen Stufe werden aus den Informationen über die detektierte Region bzw. die einzelnen Features eine möglichst diskriminative Beschreibung erzeugt, um schließlich über die Abfrage in einer Datenbank eine Identifikation zu erreichen.

Diese drei Schritte sind oft nicht strikt von einander getrennt. Sowohl Informationen, als auch verwendete Methoden werden über die verschiedenen Stufen hinweg geteilt und wiederverwendet. Speziell Detektion und *Facial Feature* Extraktion geschehen häufig parallel in einem Schritt. Durch die Betrachtung als drei separate Disziplinen ergibt sich jedoch die Möglichkeit der unabhängigen Evaluierung, Weiterentwicklung und Anwendung in anderen Gebieten.

Anwendungsgebiete

Den Schwerpunkt der Anwendungsgebiete für *Face Recognition* stellt die Sicherheitstechnik dar. Dabei werden biometrische Gesichtsmerkmale genutzt um im Wesentlichen zwei verschiedene Fragestellungen zu bedienen: Identifikation und Verifikation [Zhao et al., 2003]. Bei der Identifikation besteht das Ziel darin, einer Detektion eines Gesichtes auf einem Bild eine dem System bekannt Identität zuzuordnen. Im Unterschied dazu ist bei der Verifikation eine Vermutung über die Identität der Person schon gegeben. Es geht also darum, mit hoher Sicherheit zu bestimmen, ob es sich tatsächlich um diese Person handelt. Anwendungen finden sich in der Absicherung von Informationssystemen (Ersatz für Passwörter), Zugangskontrollen (Gebäude) oder auch Grenzkontrollen.

Ein wesentlicher Vorteil der Gesichtserkennung ist die relativ hohe soziale Akzeptanz im Vergleich zu anderen biometrischen Methoden, die zum Teil Resultate mit höherer Treffergenauigkeit (z.B. Fingerabdruck Analyse) liefern, jedoch bei unkollaborativen Personen ungeeignet sind. Dabei stellt die *Face Recognition* vergleichsweise niedrigere Anforderungen an die benötigte Hardware, als zum Beispiel Iris-Scans, wodurch sich Gesichtserkennungssysteme meist günstiger realisieren lassen. Insgesamt stellt somit *Face Recognition* einen guten Kompromiss zwischen Verlässlichkeit, sozialer Akzeptanz und Kostenaufwand dar [Abate et al., 2007].

Außerhalb der Sicherheitstechnik wird *Face Recognition* u.a. in der *Human Computer Interaction* (Spielkonsolen, Unterhaltungselektronik, *Proactive Computing*) und bei Multimedia Management (z.B.: iPhoto, Picasa) eingesetzt [Huang et al., 2005].

Herausforderungen

Die Schwierigkeiten bei der Erstellung eines Gesichtserkennungssystems decken sich zunächst mit jenen die bei jedem Versuch visueller Objekterkennung auftreten: Das Aussehen eines zu detektierenden und zu erkennenden Objekts verändert sich unter dem Einfluss der Beleuchtung (*Illumination*) und des Betrachtungswinkels (*Pose*). Zudem ist jedes Abbild mit Rauschen behaftet und Teile des Zielobjekts können verdeckt sein (*Occlusion*). Zusätzlich ist ein menschliches Gesicht kein starrer/rigider Körper, je nach Mimik verändert sich, zu einem gewissen Grad, sowohl das Aussehen einzelner Gesichtsteile, als auch die Proportionen untereinander und damit der Gesamteindruck. Ein weiterer Aspekt dem in der Forschung erst kürzlich Aufmerksamkeit zugekommen ist, ist der natürliche Alterungsprozess (*Aging*), der das menschliche Gesicht im Laufe der Zeit verändert.

Kategorisierung

In den letzten zwei Jahrzehnten wurde eine stetig und zuletzt rasant wachsende Anzahl an verschiedenen Ansätzen und Methoden zur *Face Recognition* entwickelt. Anfangs gab es das Bestreben die *Face Recognition* als allgemeines Mustererkennungsproblem zu betrachten und zu lösen. Relativ schnell stellte sich heraus, dass für effektive und akkurate Methoden spezifisches Domänenwissen über das Aussehen des menschlichen Gesichtes unabdingbar ist.

Laut [Zhao et al., 2003] lassen sie sich grob in folgende Kategorien einteilen: Holistisch, (*Facial*) *Feature-based* und Hybrid. Holistische Ansätze betrachten den Bildbereich auf dem das Gesicht abgebildet ist als ganzes, ohne sich auf spezielle Subbereiche zu konzentrieren. Prominente Vertreter aus dieser Kategorie sind *Eigenfaces* [Kirby & Sirovich, 1990, Turk & Pentland, 1991, Craw & Cameron, 1992] und *Fisherfaces* [Belhumeur et al., 1997, Swets & Weng, 1996, Zhao et al., 1998]. Hingegen wird bei *Feature-based* Ansätzen versucht einzelne charakteristische Gesichtsmerkmale (wie Augen, Nase, Mund, ...) im Bild zu erkennen. Frühe Ansätze wie [Kanade, 1973, Kelly, 1970] versuchen mit einer puristischen geometrischen Herangehensweise die Identität der Person aus der Anordnung der *Facial Features* abzuleiten. Aktuellere Arbeiten verwenden zusätzlich die *Appearance* der einzelnen *Facial Features*. Hybride Ansatze wie [Pentland et al., 1994, Lanitis et al., 1995, Penev & Atick, 1996, Huang et al., 2003] modellieren – wie auch die menschliche Wahrnehmung – das Gesicht gleichzeitig durch lokale *Features* und auch ganzheitlich.

Zusätzlich zu den genannten Kategorien können die Ansätze noch dahingehend eingeteilt werden ob sie rein auf 2D Bildern arbeiten oder (auch) die 3D Natur eines Gesichtes modellieren. Die 3D Information kann sich dabei auf verschiedenste Weise in den Ansatz eingliedern. Auch wenn nur ein 2D Testbild vorliegt kann das Kontextwissen genutzt werden, um jene Parameter zu bestimmen die die konkrete 2D Abbildung erzeugt haben. Liegt die Testszene bereits in 3D (Stereobild, Laserscan) vor, kann die *Face Recognition* auf den Vergleich der 3D Repräsentationen zurückgreifen. Eine weitere Klasse von Ansätzen nutzt sowohl die 2D als auch die 3D Information.

Übersicht über ausgewählte Methoden

In den nächsten Absätzen werden einige richtungsweisende Arbeiten aus den verschiedenen Gebieten vorgestellt und diskutiert. Der Fokus liegt hierbei auf jenen Arbeiten, die von denselben Voraussetzungen ausgehen wie sie auch in MDL gegeben sind.

Eine wichtige Gruppe, weil eine der ersten in größerem Stil erfolgreichen, bildet jene der *Subspace* Verfahren. Als bahnbrechende Arbeit sei hier als erstes die *Eigenfaces*-Methode [Kirby & Sirovich 1990; Turk & Pentland 1991] angeführt. Allgemein beruhen *Subspace* Methoden auf der Erkenntnis, dass einzelne Pixel eines Bildes das ein Gesicht enthält allgemein statistisch nicht von einander unabhängig sondern im Gegenteil sogar stark korreliert sind. Wenn man Eingabebilder der Dimension m×n als (m*n) dimensionale Vektoren betrachtet, ihnen dadurch einen Punkt in einem m*n dimensionalen Raum zuordnet, nimmt

die Gruppe aller Bilder die Gesichter enthalten einen bestimmten relativ kompakten Subbereich dieses hochdimensionalen Raumes ein. Ziel ist es nun auf mathematische Weise diesen Subraum in einen kompakten niedriger-dimensionalen Raum abzubilden, so dass in diesem anhand weniger Parameter zwischen Gesichtsbild / Nicht-Gesichtsbild bzw. zwischen den Abbildungen verschiedener Individuen unterschieden werden kann. Im Falle der *Eigenfaces* passiert diese Dimensionsreduktion mit Hilfe der *Principal Component Analysis* (PCA) [Pearson, 1901]. Die letztendliche Klassifikation erfolgt mittels *Nearest-Neighbor* Bestimmung im *Eigenspace*. Die für die PCA benötigen Vorverarbeitungschritte werden in [Craw & Cameron, 1992, Moghaddam & Pentland, 1997] untersucht. Da die Informationen, die zur Unterscheidung der einzelnen Personen dienen, nicht notwendigerweise bei der reinen PCA erhalten bleiben, greifen Methoden wie [Swets & Weng, 1996, Belhumeur et al., 1997, Zhao et al., 1998] auf *Linear Discriminant Analysis* (LDA/DFA [Fisher, 1936], *Fisherfaces* [Belhumeur et al., 1997, Swets & Weng, 1996, Zhao et al., 1998]) zurück. Weitere wichtige Vertreter von *Subspace* Methoden sind [Phillips, 1998], der sich einer *Support Vector Machine* (SVM) [Vapnik, 1995] basierten Klassifikation im *Subspace* widmet, und [Bartlett et al., 1998] der auf *Independent Component Analysis* (ICA) aufbaut. Einen sequentiellen PCA basierten Ansatz zur *Face Pose Estimation* und *Recognition* verfolgt [Liu, 2004]. Die *Pose Estimation* erfolgt mittels N*earest Neighbor to the Mean Classifier* (NNC) im niedrig dimensionalen PCA Raum. Dabei wird anhand des Abstandes zu den Mittelwertbildern für die jeweilige Pose-Klasse entschieden. Für die *Face Recognition* wird eine *Kernel PCA*, innerhalb der Pose-Klasse, auf die Gabor gefilterten Bilder angewandt, wobei auf eine Filterbank von insgesamt 40 Gabor Kernen zurückgegriffen wird. Eine weitere Möglichkeit nicht-linearer Klassifikation bieten Neuronale Netzwerke. [Lin et. al., 1997] präsentieren das Konzept der *Probabilistic Decision Based Neural Networks* (PDBNN). Für jede der zu erkennenden Personen wird dabei ein eigenes Subnetz trainiert, die Klassifikation erfolgt anhand des Subnetzes mit dem maximalen *Response*.

Einer der erfolgreichsten Vertreter aus dem Bereich der reinen *Facial Feature-based* Methoden ist *Elastic Bunch Graph Matching (EBGM)* [Okada et al. 1998; Wiskott et al. 1997]. Basierend auf *Dynamic Link Architecture* (DLA) [Lades et al. 1993] werden hier Gabor-Filter zur Erkennung, Extraktion und Beschreibung wesentlicher Gesichtsmerkmale eingesetzt. *Sets* von Gabor-Filtern (*Jets*) sind dabei jeweils einem Knoten in einem Graphen zugeordnet, der eine rohe geometrische Darstellung eines Gesichts beschreibt. Das System wird trainiert, indem auf den ersten Lerndaten die Positionen der Knoten im Bild manuell markiert werden. Die Auswertung der Gabor-*Jets* liefert Parameter, die die jeweiligen Gesichtsmerkmale beschreiben und gelernt werden. Schon nach einigen dutzend Trainingsbeispielen haben die Gabor-*Jets* eine Repräsentation gelernt die ausreicht, um auf weiteren Lerndaten die Position der Features über eine Rastersuche und geometrische Validierung automatisch zu finden. Über den gelernten EBG erhält man für ein neu zu klassifizierendes Input-Bild demnach Position und Beschreibung der Geschichtsmerkmale. Um den aufwendigen Matching-Prozess der dabei nötig ist zu vermeiden wird in [Perronnin & Dugelay, 2003] ein deformierbares Modell auf Basis einer 2D Erweiterung eines 1D-HMM (*Hidden Markov Model*) vorgestellt.

Die Entwicklung hybrider Ansätze begründet eine modulare, *Facial-Feature*-basierte Erweiterung von *Eigenfaces* [Pentland et al., 1994]. Dabei werden zusätzlich zum holistischen *Eigenface Eigenfeatures* wie *Eigeneyes*, *Eigenmouth* etc. modelliert.

Mit *Active Shape Models* (ASM) wird in [Lanitis et al., 1995, Cootes et al., 1995] ein Ansatz vorgestellt, der ein parametrisches Modell der Geometrie des Gesichtes verwendet. Die *Shape* Parameter werden dazu verwendet die Gesichtstextur normalisiert (*shape-free*) zu beschreiben, lokale Merkmale besser zu lokalisieren und somit die Klassifikation zu vereinfachen. Außerdem können die gefundenen *Shape* Parameter auch direkt zur Klassifikation hinzugezogen werden.

Active Appearance Models [Edwards et al., 1998, Cootes et al., 2000] erweitern ASMs, indem sie die Suche nach den korrekten Parametern sowohl für das *Shape*-Modells als auch für das *Shape*-neutralen Textur-Modells (z.B. *Eigenfaces*) zu einem einheitlichen iterativen Prozess kombinieren. Zur Klassifikation wird hier mittels LDA versucht einen *Subspace* zu erzeugen, in dem die einzelnen Individuen möglichst gut voneinander trennbar sind (*Identity Subspace*) und orthogonal dazu ein Residuen-*Subspace,* der nur die Variationen innerhalb des Aussehens einer Person aufspannt. In den genannten Arbeiten wird AAM größtenteils für *Face-Tracking* verwendet, ohne dabei konkrete Resultate für *Recognition* anzugeben. Neuere Arbeiten wie [Faggian et al., 2006] verwenden AAM vorwiegend als Vorverarbeitungsschritt zur *Pose Estimation/Refinement* und Segmentierung des Gesichtsbereichs, um in weiterer

Folge konventionelle Erkennungsmethoden auf *Shape*-neutralen Gesichtsbildern anwenden zu können. [Kim et al., 2007] verwenden ein AAM im Kontext eines von EBGM abgeleiteten Systems um eine bessere Initialisierung für die Suche der Gabor-*Feature*-Punkte zu erzielen.

Eine logische Erweiterung von 2D parametrischen, modellbasierten *Face Recognition* Ansätzen ist die Berücksichtigung von 3D Information zur Modellbildung. Dabei werden die 3D parametrischen Modelle in das 2D Eingangsbild eingepasst, um so Faktoren wie Pose und Illumination besser zu erfassen. Der Ansatz von [Blanz & Vetter, 2002] beruht auf einem generischen *3D Morphable Model* (3DMM). Ein 3DMM ist ein deformierbares Kopfmodell das Textur- und *Shape*-Parameter unabhängig voneinander, durch eine jeweilige Vektorbasis, modelliert. Die Vektorbasis wird dabei durch die getrennte PCA-Analyse (*Shape*,Textur) einer Datenbank von 3D *Scans* erhalten. Die Identifikation von Gesichtern erfolgt, unabhängig von separat modellierten Einflüssen wie *Pose* und *Illumination*, durch die aus einem einzelnen 2D Bild wiedererlangten Modellparameter. Eine wesentliche Einschränkung dieses Ansatzes ist das Konvergenzverhalten, das eine gute Initialisierung für den komplexen Optimierungsprozess, der 22 Parameter berücksichtigt, erfordert und außerdem sehr rechenaufwändig ist. [Smet et. al., 2006] widmet sich dem Problem der Robustheit gegenüber partiellen Verdeckungen. In einem *Generalized Expectation Maximization* (GEM) Prozess wird die Schätzung der Modellparameter vereint mit der Berechnung einer *Visibility Map,* repräsentiert durch ein *Markov Random Field* (MRF).

Zusammenfassung und Diskussion

Der Überblick über die *Face Recognition* Literatur legt die Schlussfolgerung nahe, dass sich noch keine einheitliche Stoßrichtung abzeichnet. Gerade in den letzten Jahren wurden viele Methoden entwickelt von denen jede versucht spezifische Probleme zu lösen jedoch noch keine allen Anforderungen gerecht wird und problemlos universell einsetzbar ist.

Um Vergleiche zwischen den divergierenden Ansätzen zu ziehen wurden wiederholt Evaluierungen auf gemeinsamen Datensätzen durchgeführt. Einerseits um den Fortschritt der einzelnen Methoden zu beurteilen, andererseits um durch neue Testszenarien Herausforderungen aufzuzeigen. Die erste größere Bestrebung eine gemeinsame Evaluierungsplattform zu schaffen führte zur FERET (FacE REcogntion Technology) Datenbank. Einige der getesteten Algorithmen haben gezeigt, dass das Problem der Face Recognition unter kontrollierten Bedingungen zufriedenstellend lösbar ist. Deutliche Herausforderungen bilden Einflüsse wie *Pose*, *Illumination*, *Clutter* und *Occlusion*, die im speziellen bei Outdoor-Szenarien auftreten. Weitere Evaluierungen wie die *Face Recognition Grand Challenge* (FRGC) oder der *Face Recognition Vendor* Test (FRVT) legten speziell einen Schwerpunkt auf diese Herausforderungen.

Der FRVT wurde bisher dreimal durchgeführt (2000,2002,2006). Gedacht ist FRVT vorranging als Benchmark für kommerziell verfügbare (proprietäre) Systeme, weniger zur Evaluierung und Weiterentwicklung neuer Algorithmen. Nichts desto trotz gibt FRVT Aufschluss darüber, was zurzeit mit aktuellen Methoden in der *Face Recognition* möglich ist und welche Bereiche noch nicht als gelöst betrachtet werden können. Beim FRVT 2002 hat sich gezeigt, dass die teilnehmenden Systeme mit Aufnahmen unter relativ kontrollierten Bedingungen gut zurecht kommen. Die wesentlichen Problemfelder die aufgezeigt werden konnten sind: Aufnahmen unter gänzlich unkontrollierten Bedingungen (Außenaufnahmen); die Zeitspanne zwischen den einzelnen Aufnahmen (Alterung, die Wiedererkennungsrate sinkt in etwa um 5% pro Jahr); Ältere Menschen werden leichter erkannt als Jüngere; Frauen werden in etwa um 6-9% schlechter erkannt als Männer. Zu den bekannten und nach wie vor nicht zur Gänze gelösten Problemen großer Änderungen in Pose und Beleuchtung und Verdeckungen sind als noch demografische Spezifika gekommen [Phillips et al., 2005]. Weitere Fortschritte im Bereich der Aufnahmen unter unkontrollierten Bedingungen wurden beim FRVT 2006 festgestellt. Die Wiedererkennungsraten in diesem Bereich sind vergleichbar mit jenen des FRVT 2002 für rein kontrollierte Bedingungen. Weiters wurde erstmals gezeigt, dass unter gewissen Einflüssen wie Beleuchtungsänderungen gewisse Algorithmen die Erkennungsfähigkeiten des Menschen übertreffen können [Phillips et al., 2009].

Bezogen auf die oben vorgestellten Methoden verspricht man sich von modellbasierten Ansätzen wie AAM und 3DMMs mehr Invarianz gegenüber den angesprochenen Herausforderungen wie *Pose* und *Illumination* Einflüssen. Diese werfen aber andere Probleme auf wie etwa Konvergenz oder Laufzeit. Daher kann man davon ausgehen, dass keine der existierenden Methoden allen Anforderungen und Herausforderungen gleichzeitig, zufriedenstellend gerecht wird [Abate et al. 2007]. Ein Überblick über die Resultate der einzelnen Methoden auf den jeweiligen Testdatensätzen wird in [Abate et al., 2007] angeführt.

Literaturverzeichnis

[Abate et al., 2007] Abate, A. F., Nappi, M., Riccio, D., und Sabatino, G. (2007). 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906.

[Bartlett et al., 1998] Bartlett, M. S., Lades, M. H., Sejnowski, T. J., Rogowitz, B. E., und Pappas, T. N. (1998). Independent component representations for face recognition. In *Proc. SPIE: Human Vision and Electronic Imaging III*, Band 3299, Seiten 528–539.

[Belhumeur et al., 1997] Belhumeur, P. N., Hespanha, J. P., und Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):711–720.

[Biederman & Kalocsai, 1998] Biederman, I. und Kalocsai, P. (1998). Neural and psychophysical analysis of object and face recognition. In Wechsler, H., Phillips, P., Bruce, V., Soulie, F., und Huang, T., Editoren, *Face Recognition: From Theory to Applications, NATO ASI Series F*, Band 163, Seiten 3–25. Springer.

[Blanz et al., 2002] Blanz, V., Romdhani, S., und Vetter, T. (2002). Face identification across different poses and illuminations with a 3D morphable model. In *Proc. IEEE Intern. Conf. on Automatic Face and Gesture Recognition*, Seiten 192–197.

[Cootes et al., 2001] Cootes, T., Edwards, G., und Taylor, C. (2001). Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685.

[Cootes et al., 1995] Cootes, T. F., Taylor, C. J., Cooper, D. H., und Graham, J. (1995). Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1):38–59.

[Cootes et al., 2000] Cootes, T. F., Walker, K., und Taylor, C. J. (2000). View-Based active appearance models. In *Proc. IEEE Intern. Conf. on Automatic Face and Gesture Recognition*, Seite 227.

[Craw & Cameron, 1992] Craw, I. und Cameron, P. (1992). Face recognition by computer. In *Proc. British Machine Vision Conf.*, Seiten 498–507.

[Distasi et al., 2003] Distasi, R., Nappi, M., und Tucci, M. (2003). FIRE: fractal indexing with robust extensions for image databases. *IEEE Trans. on Image Processing*, 12(3):373–384.

[Edwards et al., 1998] Edwards, G., Taylor, C., und Cootes, T. (1998). Learning to identify and track faces in image sequences. In *Proc. IEEE Intern. Conf. on Automatic Face and Gesture Recognition*, Seiten 260–265.

[Ellis, 1986] Ellis, H. (1986). Introduction to aspects of face processing: Ten questions in need of answers. *Aspects of Face Processing*.

[Faggian et al., 2006] Faggian, N., Paplinski, A., und Chin, T. (2006). Face recognition from video using active appearance model segmentation. In *Proc. Intern. Conf. on Pattern Recognition*, Band 1, Seiten 287–290.

[Fisher, 1936] Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

[Gao & Leung, 2002] Gao, Y. und Leung, M. (2002). Face recognition using line edge map. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(6):764–779.

[Gauthier et al., 1999] Gauthier, I., Behrmann, M., und Tarr, M. J. (1999). Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience*, 11(4):349–370.

[Gauthier & Logothetis, 2000] Gauthier, I. und Logothetis, N. K. (2000). Is face recognition not so unique after all? *Cognitive Neuropsychology*, 17:125–142.

[Huang et al., 2003] Huang, J., Heisele, B., und Blanz, V. (2003). Component-Based face recognition with 3D morphable models. In *Audio- and Video-Based Biometric Person Authentication*, Seiten 27–34. Springer Berlin / Heidelberg.

[Huang et al., 2005] Huang, T., Xiong, Z., und Zhang, Z. (2005). Face recognition applications. In Li, S. Z. und Jain, A. K., Editoren, *Handbook of Face Recognition*, Seiten 371–390. Springer.

[Jain et al., 2004] Jain, A., Ross, A., und Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(1):4–20.

[Kanade, 1977] Kanade, T. (1977). Computer recognition of human faces. *Interdisciplinary Systems Research*, 47.

[Kelly, 1971] Kelly, M. D. (1971). *Visual identification of people by computer*. Doktorarbeit, Stanford University.

[Kim et al., 2007] Kim, S., Chung, S., Jung, S., Jeon, S., Kim, J., und Cho, S. (2007). Robust face recognition using AAM and Gabor features. In *Proc. of the World Academy of Science, Engineering and Technology*, Band 33, Seiten 117–121.

[Kirby & Sirovich, 1990] Kirby, M. und Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1):103–108.

[Lades et al., 1993] Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R., und Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311.

[Lanitis et al., 1995] Lanitis, A., Taylor, C., und Cootes, T. (1995). Automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–401.

[Lin et al., 1997] Lin, S., Kung, S., und Lin, L. (1997). Face recognition/detection by probabilistic decision-based neural network. *IEEE Trans. on Neural Networks*, 8(1):114–132.

[Liu, 2004] Liu, C. (2004). Gabor-based kernel PCA with fractional power polynomial models for face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(5):572–581.

[Lowe, 1999] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc. IEEE Intern. Conf. on Computer Vision*, Band 2, Seiten 1150–1157.

[Moghaddam & Pentland, 1997] Moghaddam, B. und Pentland, A. (1997). Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):696–710.

[Palanivel & Yegnanarayana, 2008] Palanivel, S. und Yegnanarayana, B. (2008). Multimodal person authentication using speech, face and visual speech. *Computer Vision and Image Understanding*, 109(1):44–55.

[Pearson, 1901] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572.

[Penev & Atick, 1996] Penev, P. S. und Atick, J. J. (1996). Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477 – 500.

[Pentland et al., 1994] Pentland, A., Moghaddam, B., und Starner, T. (1994). View-based and modular eigenspaces for face recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 84–91.

[Perronnin & Dugelay, 2003] Perronnin, F. und Dugelay, J. (2003). An introduction to biometrics and face recognition. In Proc. *IMAGE:* First Intern. Workshop on *Learning, Understanding, Information Retrieval, Medical*, Seiten 1–20.

[Phillips et al., 2009] Phillips, P., Scruggs, W., O'Toole, A., Flynn, P., Bowyer, K., Schott, C., und Sharpe, M. (2009). FRVT 2006 and ICE 2006 Large-Scale experimental results. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* Zur Veröffentlichung akzeptiert.

[Phillips, 1998] Phillips, P. J. (1998). Support vector machines applied to face recognition. In *Advances in Neural Information Processing Systems 11*, Seiten 803–809. MIT Press.

[Phillips et al., 2005] Phillips, P. J., Grother, P., und Micheals, R. (2005). Evaluation methods in face recognition. In Li, S. Z. und Jain, A. K., Editoren, *Handbook of Face Recognition*, Seiten 329–348. Springer.

[Schneiderman & Kanade, 2000] Schneiderman, H. und Kanade, T. (2000). A statistical method for 3D object detection applied to faces and cars. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Band 1, Seiten 746–751.

[Smet et al., 2006] Smet, M. D., Fransens, R., und Gool, L. V. (2006). A generalized EM approach for 3D model based face recognition under occlusions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Band 2, Seiten 1423–1430.

[Steffens & Okada, 1998] Steffens, J. und Okada, K. (1998). The Bochum/USC face recognition system and how it fared in the FERET phase III test. In Wechsler, H., Phillips, P., Bruce, V., Soulie, F., und Huang, T., Editoren, *Face Recognition: From Theory to Applications, NATO ASI Series F*, Band 163, Seiten 186–205. Springer.

[Swets & Weng, 1996] Swets, D. und Weng, J. (1996). Using discriminant eigenfeatures for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(8):831–836.

[Turk & Pentland, 1991] Turk, M. und Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

[Vapnik, 1995] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.

[Wiskott et al., 1997] Wiskott, L., Fellous, J., Kuiger, N., und von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):775–779.

[Zhao et al., 1998] Zhao, W., Chellappa, R., und Krishnaswamy, A. (1998). Discriminant analysis of principal components for face recognition. In *Proc. IEEE Intern. Conf. on Automatic Face and Gesture Recognition*, Seiten 336–341.

[Zhao et al., 2003] Zhao, W., Chellappa, R., Phillips, P. J., und Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458.

Face Tracking und Recognition in Videos

Die ersten Methoden, die sich mit *Face Tracking* und *Recognition* in Videos beschäftigten betrachten die Videosequenzen einfach als Serie von Einzelbildern und wenden existierende Verfahren für Detektion und Erkennung von Gesichtern auf statischen Bildern an, ohne den zeitlichen Bezug der Bilder zu einander zu berücksichtigen.

Eine Möglichkeit die Bilder nicht ganz separat zu betrachten, sondern miteinander lose in Relation zu bringen, ist die Ergebnisse aus den Einzelframes in einem probabilistischen *Voting* zusammenzuführen, um eine stabilere Identifikation zu erreichen [Gong et al., 2000, McKenna & Gong, 1998]. Ein erster Schritt zur Ausnützung der temporalen Information bildet die Kombination von *Tracking* und Recognition in einen sequentiellen Prozess. Dabei wird die temporale Information nur im *Tracking* berücksichtigt, die *Recogntion* arbeitet weiterhin auf Einzelframes und nur dann wenn ein gewisses Qualitätskriterium der *Face Detection* (wie etwa Größe oder *Pose*) erfüllt ist [Choudhury et al., 1998, Li & Chellappa, 2001, Dedeoglu et al., 2007]. Die verwendeten *Tracking* Ansätze können dabei modellbasiert mit Domänenwissen arbeiten, wie etwa [Cootes et al., 2001] die auf AAMs zurückgreifen.

Weiterführend wird versucht die temporale Information sowohl im *Tracking* als auch in der *Recognition* mit einzubeziehen [Zhou et al., 2003, Chellappa & Zhou, 2005]. [Li et al., 2001, Li et al., 2003] extrahieren aus den Trainingsvideos framebasiert die einzelnen Gesichtsposen, um pro Person eine *Identity Surface* zu erzeugen, die in weiterer Folge zur Recognition verwendet wird. Die *Identity Surface* ordnet jedem Subjekt unter einer gewissen *Pose* (*yaw*, *tilt*) eine bestimmte *Appearance* zu. Während eine einzelne Beobachtung im Parameteraum oft nicht eindeutig einem *Identity Surface* zugeordnet werden kann, ergibt eine Serie von Beobachtungen eine Trajektorie die deutlich mehr Information, zur Identifikation, bietet. [Xiao et. al., 2004] stellen eine Methode vor, die es erlaubt aus einer Sequenz von Bildern das explizite 3D Modell einer Person abzuleiten, das in weiterer Folge für eine *Pose Estimation* und verbesserte *Recognition* verwendet werden kann. Dabei wird das 3D Modell aus den AAM Parametern rekonstruiert, die beim *Fitting* Prozess, eines getrackten Gesichts, entstehen. Durch die Möglichkeit 3D *Shape* und *Pose* zu berechnen vereint der Ansatz die Vorteile eines 3DMMs mit der Echtzeitfähigkeit eines 2D AAMs. Durch eine Beschränkung des 2D AAM Parameterraumes, basierend auf dem 3D Modell, führt der Ansatz sogar zu schnellerer Konvergenz als das rein 2D basierte AAM.

Die Kombination von temporalen und räumlichen Informationen für Face Tracking und Recognition führt einerseits zu Leistungsgewinnen da nicht jedes Bild prozessiert werden muss andererseits kann die geringere räumliche Auflösung durch temporale Information kompensiert werden. Im Bereich der Recognition wird Methoden die gleichzeitig temporale und räumliche Information einbeziehen durchaus mehr Potential eingeräumt als diese bisher zeigen konnten.

Literaturverzeichnis

[Chellappa & Zhou, 2005] Chellappa, R. und Zhou, S. K. (2005). Face tracking and recognition from video. In Li, S. Z. und Jain, A. K., Editoren, *Handbook of Face Recognition*, Seiten 169–192. Springer.

[Choudhury et al., 1998] Choudhury, T., Clarkson, B., Jebara, T., und Pentl, A. (1998). Multimodal person recognition using unconstrained audio and video. In *Proceedings International Conference on Audio- and Video-Based Person Authentication*, Seiten 176–181.

[Cootes et al., 2001] Cootes, T., Edwards, G., und Taylor, C. (2001). Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685.

[Dedeoglu et al., 2007] Dedeoglu, G., Kanade, T., und Baker, S. (2007). The asymmetry of image registration and its application to face tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(5):807–823.

[Gong et al., 2000] Gong, S., McKenna, S. J., und Psarrou, A. (2000). *Dynamic Vision: From Images to Face Recognition*. Imperial College Press.

[Li & Chellappa, 2001] Li, B. und Chellappa, R. (2001). Face verification through tracking facial features. *Journal of the Optical Society of America A*, 18(12):2969–2981.

[Li et al., 2001] Li, Y., Gong, S., und Liddell, H. (2001). Constructing facial identity surfaces in a nonlinear discriminating space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Band 2, Seiten 258–263.

[Li et al., 2003] Li, Y., Gong, S., und Liddell, H. (2003). Constructing facial identity surfaces for recognition. *Intern. Journal of Computer Vision*, 53(1):71–92.

[McKenna & Gong, 1998] McKenna, S. und Gong, S. (1998). Recognizing moving faces. In Wechsler, H., Phillips, P., Bruce, V., Soulie, F., und Huang, T., Editoren, *Face Recognition: From Theory to Applications, NATO ASI Series F*, Band 163, Seiten 578–588. Springer.

[Xiao et al., 2004] Xiao, J., Baker, S., Matthews, I., und Kanade, T. (2004). Real-time combined 2D+3D active appearance models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Band 2, Seiten 535–542.

[Zhou et al., 2003] Zhou, S., Krueger, V., und Chellappa, R. (2003). Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91(1-2):214–245.

Map Detection

Eine der (zentralen) Fragestellungen im Projekt MDL ist es, Nachrichtensendungen automatisiert dahin gehend zu analysieren mit welchen Regionen der Erde die einzelnen Berichte inhaltlich in Bezug stehen. Dabei liegt ein Hauptaugenmerk in der visuellen Verarbeitung darin eingeblendete Karten zu detektieren und in weiterer Folge daraus Informationen zu extrahieren; unabhängig von den Analysemitteln des Audiostreams.

Das Thema der Karten-Detektion und Erkennung war bislang noch kaum Gegenstand intensiver Forschung. Ein grundlegendes Problem für die automatisierte Verarbeitung mit Mitteln der Computer Vision stellt die hohe Variabilität der Darstellungs- und Erscheinungsformen der semantischen Kategorie „Karte“ dar. Was man als Mensch als Karte einordnen würde, reicht von einer abstrakten, schematisch reduzierten Darstellung (etwa der Plan eines U-Bahnnetzes) bis hin zu annotierten Satellitenbildern. Die Form der Darstellung variiert in der Perspektive von orthografischen Projektionen, über Vogelperspektive bis hin zu dreidimensionalen Reliefmodellen. Je nach Anwendungsgebiet verfügt eine Karte über unterschiedliche Annotationen (Beschriftungen und Symbole) und Kolorierungen. Je mehr von der Variabilität mit diesem System erfasst werden soll desto komplexer gestaltet sich die Problemstellung. Für eine funktionierende technische Lösung scheint es daher sinnvoll, den Typus der Karte die erkannt werden soll im Vorfeld, anhand der charakteristischen visuellen Eigenschaften, genau zu definieren.

In einem ersten Schritt kann das Problem darauf reduziert werden, Bilder in Karten und nicht Karten zu klassifizieren. Dieser Fragestellung widmen sich [Michelson et al., 2008]. Um die Variabilität der verwendeten Testdaten zu modellieren wird vorgeschlagen auf *Content Based Image Retrieval* (CBIR) zurückzugreifen. Im Unterschied zu konventionellen Klassifikationsmethoden aus dem Maschinellem Lernen. Dabei wird ein Testbild anhand eines Ähnlichkeitsmaßes, generiert aus *Water Filling Features* [Zhou et al., 1999], mit einer großen Datenbank verglichen. Die Wahl der *Features* zeigt eindeutig eine Fokussierung auf Karten mit klaren Linienzügen und Kantenstrukturen (wie etwa Stadtpläne) und schränkt damit den Anwendungsbereich ein. Die Klassifikation erfolgt mittels einer *k-Nearest Neighbor* Mehrheitsentscheidung. Dabei wird die Menge der Kartentypen rein durch die Trainingsdaten implizit repräsentiert. Bedingt durch die Unterschiede zwischen den einzelnen Kartentypen, die vorab nicht bekannt sind und die daraus resultierende Inhomogenität der gemeinsamen Klasse aller Kartentypen wird die bessere Erkennungsrate gegenüber SVM basierten Ansätzen argumentiert. Die Autoren argumentieren, dass sich die bessere Erkennungsrate gegenüber dem SVM basierten Ansatz auf zwei Faktoren zurückführen lässt: Die Menge aller Karten lässt sich vorab schwer einzelnen Klassen zuordnen und die Gesamtheit aller Karten bildet im Feature Space keine homogenen Subraum.

Das Problem der Kartendetektion kann als wissenschaftliches Neuland angesehen werden. Da folglich wenig Literatur zur Verfügung steht existiert auch noch keine klare Analyse und Abgrenzung der Aufgabenstellung bzw. Vorschläge zu deren Lösung. Eine notwendige Vorbedingung ist sicher Datenmaterial zu sammeln und auf diesem eine genaue Analyse welche Arten von Karten und Formen der Abbildung es zu erkennen gilt. Vergleicht man die visuellen Charakteristika etwa von einem annotierten Satellitenbild mit denen einer abstrakten Landkarte, stellt sich die Frage inwieweit die jeweiligen Abstraktionsebenen im Sinne der Objekterkennung als eine gemeinsame Klasse betrachtet werden und demnach auch gemeinsam modelliert werden können, oder ob doch für jeden Typ eine eigene Detektionsmethode entwickelt werden muss. [Michelson et al., 2008] argumentiert zwar mit einem generischen Ansatz die Variabilität abzudecken, aus der Wahl der verwendeten Features wird jedoch klar, dass nur stark abstrahierte Karten mit klaren Linienzügen und Kantenstrukturen (wie Stadtpläne, Liniennetze öffentlicher Verkehrsbetriebe oder Infrastrukturkarten) abgedeckt werden. In weiterer Folge wird es erforderlich sein allgemeine Algorithmen zur Segmentierung und Featuregenerierung auf dieses Anwendungsgebiet anzupassen und im Kontext dieser Aufgabenstellung weiterzuentwickeln und zu evaluieren.

Literaturverzeichnis

[Michelson et al., 2008] Michelson, M., Goel, A., und Knoblock, C. A. (2008). Identifying maps on the world wide web. In *Proc. Intern. Conf. on Geographic Information Science*, Seiten 249–260. Springer.

[Zhou et al., 1999] Zhou, X. S., Rui, Y., und Huang, T. (1999). Water-filling: a novel way for image structural feature extraction. In *Proc. IEEE Intern. Conf. on Image Processing*, Band 2, Seiten 570–574.

Appendix

Datenbanken

FERET

Die FERET Datenbank enthält 14.126 Bilder von 1.199 Personen. Die Aufnahmen sind Stile von Passbildern mit Variation der Pose vom linken bis zum rechten Profil. Die Bilder sind unter kontrollierten Bedingungen entstanden. Bei einer Untermenge der Personen wurden weiters Duplikate der Bilder nach einer gewissen Zeitspanne angefertigt. Abbildung 3 zeigt ein Beispiel aus der Datenbank.

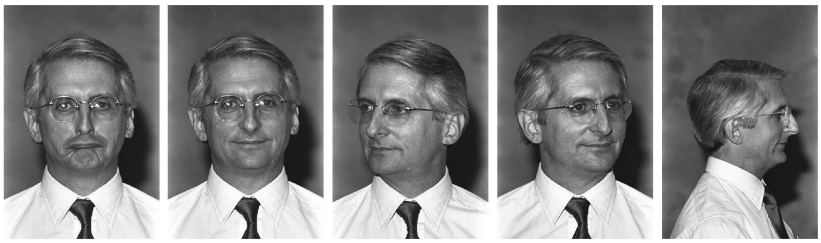

Abbildung 3 FERET [Phillips et al., 2000].

CMU-MIT Frontal

Die CMU-MIT Datenbank für frontale Gesichtsdetektion enthält 130 Bilder auf denen 507 Gesichter annotiert sind. Die verwendeten Bilder weisen deutliche Variationen (Beleuchtung, Skalierung, Anzahl der Gesichter im Bild) auf, siehe auch Abbildung 4.

Abbildung 4 CMU-MIT Frontal [Rowley et al., 1996].

CMU Profil

Die CMU Datenbank für nicht-frontale Gesichtsdetektion besteht aus 208 Bildern die 441 Gesichter zeigen, wovon 347 als Profilaufnahmen annotiert sind. Abbildung 5 zeigt einige Beispiele aus dieser.

Abbildung 5 CMU Profile [Schneiderman & Kanade, 2000].

CMU-PIE

CMU-PIE (kurz für *Pose*, *Illumination*, and *Expression*) enthält insgesamt 41.368 Bilder von 68 Personen. Von jeder Person wurden Aufnahmen in 13 verschiedenen Posen (siehe Abbildung 6), unter 43 verschiedenen Beleuchtungssituationen und mit vier verschiedenen Gesichtsausdrücken gemacht.

Abbildung 6 CMU-PIE [Sim et al., 2002].

Literaturverzeichnis

[Phillips et al., 2000] Phillips, P., Moon, H., Rizvi, S., und Rauss, P. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104.

[Rowley et al., 1996] Rowley, H., Baluja, S., und Kanade, T. (1996). Neural network-based face detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Seiten 203–208.

[Sim et al., 2002] Sim, T., Baker, S., und Bsat, M. (2002). The CMU pose, illumination, and expression (PIE) database. In *Proc. IEEE Intern Conf. on Automatic Face and Gesture Recognition*, Seiten 46–51.

[Schneiderman & Kanade, 2000] Schneiderman, H. und Kanade, T. (2000). A statistical method for 3D object detection applied to faces and cars. *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Band 1, Seiten 746–751.