Die Detektion interessanter Objekte unter Verwendung eines objektbasierten Aufmerksamkeitsmodells

Dissertation

zur Erlangung des Grades eines

D o k t o r s d e r I n g e n i e u r w i s s e n s c h a f t e n

der Technischen Universität Dortmund
an der Fakultät für Informatik

von

Fabian Naße

Dortmund

2016

Tag der mündlichen Prüfung:	7. Dezember 2016
Dekan / Dekanin:	Prof. Dr.-Ing. Gernot A. Fink
Gutachter / Gutachterinnen:	Prof. Dr.-Ing. Gernot A. Fink
	Prof. Dr. Christian Wöhler

DIE DETEKTION INTERESSANTER OBJEKTE UNTER VERWENDUNG EINES OBJEKTBASIERTEN AUFMERKSAMKEITSMODELLS

Zusammenfassung

Das visuelle System des Menschen ist in der Lage, komplexe Aufgaben, wie beispielsweise das Erkennen von Objekten und Personen, problemlos zu bewältigen. Mit dem Begriff Computer-Vision wird ein Forschungsgebiet bezeichnet, bei der die Fragestellung im Vordergrund steht, wie eine vergleichbare Leistungsfähigkeit in technischen Systemen erreicht werden kann. In dieser Dissertation wird diesbezüglich das Prinzip der visuellen Aufmerksamkeit betrachtet, dass einen wichtigen Aspekt des menschlichen Sehsystems darstellt. Es besagt, dass der bewussten Wahrnehmung ein unbewusster Prozess vorausgeht, durch den die Aufmerksamkeit selektiv auf potentiell wichtige oder interessante Sehinhalte gelenkt wird. Es handelt sich dabei um eine Strategie der effizienten Informationsverarbeitung, die ein schnelles Reagieren auf relevante Inhalte erlaubt. In diesem Zusammenhang bezeichnet der Begriff der visuellen Salienz die Eigenschaft von Sehinhalten, im Vergleich zu ihrem Umfeld hervorzustechen und deshalb Aufmerksamkeit zu stimulieren. Im Allgemeinen besteht für solche Inhalte eine vergleichsweise hohe Wahrscheinlichkeit, dass sie für das beobachtende Individuum von Interesse sind. Diese Arbeit hat das Thema der aufmerksamkeitsbasierten Objektdetektion zum Gegenstand. Motiviert wird das Thema als eine Alternative zu wissensbasierten Objektdetektionsverfahren, bei denen Klassifizierungsmodelle mittels annotierten Beispielbildern angelernt werden. Solche Verfahren sind im Allgemeinen mit einem hohen manuellen Vorbereitungsaufwand verbunden, weisen eine hohe Komplexität auf und skalieren schlecht mit der Anzahl der betrachteten Objektkategorien. Die zentrale Fragestellung dieser Arbeit ist es deshalb, ob sich Salienz als Kriterium für eine effizientere Lokalisierung von Objekten in Bildern nutzen lässt. Aufbauend auf der These, dass gerade die interessanten Objekte einer Szene visuell salient sind, soll durch einen aufmerksamkeitsbasierten Ansatz eine schnelle und aufwandsarme Detektion solcher Objekte ermöglicht werden. Es werden in dieser Arbeit zunächst wichtige Grundlagen aus den Bereichen der Mustererkennung, des maschinellen Lernens und der Bildverarbeitung erläutert. Anschließend werden klassische Strategien zur Lokalisierung von Objekten in Bildern aufgezeigt. Dabei werden Vor- und Nachteile verschiedener Lokalisierungsstrategien im Hinblick auf den aufmerksamkeitsbasierten Ansatz betrachtet. Im Anschluss daran werden grundlegende Konzepte sowie einflussreiche Theorien und Modelle zur visuellen Aufmerksamkeit des Menschen aufgezeigt. Hieran schließt sich eine Betrachtung mathematischer Aufmerksamkeitsmodelle aus der Literatur an. Aufbauend darauf wird ein eigenes Aufmerksamkeitsmodell vorgeschlagen, dass Objektvorschläge ermittelt und anhand ihrer Salienz bewertet. Zwecks einer generischen Anwendbarkeit wird dabei ein rein datengetriebener

Ansatz favorisiert, bei dem bewusst auf die Verwendung problemspezifischen Vorwissens verzichtet wird. Das Verfahren wird schließlich auf einem schwierigen Benchmark evaluiert. Dabei werden durch Vergleiche mit anderen Modellen aus der Literatur die Vorteile der vorgeschlagenen Methoden hervorgehoben. Des Weiteren wird bei der Betrachtung der Ergebnisse gezeigt, dass Salienz ein wichtiges Kriterium bei der generischen Lokalisierung von Objekten in komplexen Bildern darstellt.

INHALTSVERZEICHNIS

1 EINLEITUNG

Das visuelle System des Menschen ist sehr leistungsfähig (Jenkin und Harris [1], S. 3 f.). Es ist in der Lage, komplexe Aufgaben wie beispielsweise das Erkennen von Objekten und Personen problemlos zu bewältigen. Einen wichtigen Aspekt des Sehsystems stellt das Prinzip der visuellen Aufmerksamkeit (Jenkin und Harris [1], S. 1 ff.) dar, welches eine bedeutende Rolle bei der effizienten Auswertung von Sehinhalten spielt. Aufmerksamkeit bedeutet, dass die bewusste Wahrnehmung auf potentiell wichtige Sehinhalte gelenkt wird. Der entsprechende Prozess wird *selektive Aufmerksamkeit* (Jenkin und Harris [1], S. 2) genannt. Er bestimmt, welche der durch die Sinnesorgane aufgenommenen Signale in das Bewusstsein vordringen und welche nicht. Das bedeutet, der bewussten Wahrnehmung geht ein unbewusster Verarbeitungsprozess voraus, der die Aufmerksamkeit des beobachtenden Individuums auf bestimmte Inhalte lenkt und andere Inhalte vernachlässigt. Es handelt sich dabei um eine Strategie der effizienten Informationsverarbeitung, die ein schnelles Reagieren auf wesentliche Inhalte erlauben soll. Wichtig im Zusammenhang mit der visuellen Aufmerksamkeit ist der Begriff der visuellen *Salienz* (Jenkin und Harris [1], S. 8). Aufmerksamkeit setzt voraus, dass Signale bestimmte Eigenschaften aufweisen, um bewusst wahrgenommen zu werden. Man spricht in diesem Zusammenhang von *Reizen* oder auch *Stimuli* (Einzahl: Stimulus; Jenkin und Harris [1], S. 8). Wie oben bereits angesprochen, wirken Inhalte insbesondere dann stimulierend, wenn sie salient sind. Salienz bezeichnet die Eigenschaft eines Reizes, sich von der Mehrheit der zeitnah oder parallel auftretenden Reize aufgrund bestimmter grundlegender Eigenschaften zu unterscheiden. Im Falle der visuellen Wahrnehmung kann es sich um auffällige Farben oder Formen handeln, die einen bestimmten Bereich innerhalb des Sichtfeldes hervorheben. Im Allgemeinen besteht für solche Bereiche eine höhere Wahrscheinlichkeit, dass sie für das beobachtende Individuum von Interesse sind, als für unauffälligere Bereiche. Der Prozess der selektiven Aufmerksamkeit läuft kontinuierlich und mit einer hohen Geschwindigkeit ab. Eine tiefere Interpretation der Inhalte findet hingegen nicht statt und ist der Phase der bewussten Wahrnehmung vorbehalten. Das effiziente Lenken auf wichtige Inhalte macht das Prinzip der selektiven Aufmerksamkeit für technische Anwendungen interessant. In dieser Arbeit geht es um das Thema der objektbasierten Aufmerksamkeit (Jenkin und Harris [1], S. 9). Bei dieser wird davon ausgegangen, dass Objekte im Sichtfeld des Betrachters um Aufmerksamkeit konkurrieren. Dieses Prinzip erlaubt es, die potentiell interessanten Objekte einer Szene auszumachen. Ein Objekt gilt dabei als interessant, wenn es einen relevanten Beitrag zum Ver-

ständnis des Szeneninhalts liefert. Die Detektion solcher Objekte stellt einen wichtigen Aspekt bei der Entwicklung technischer Systeme dar, die mit ihrer Umgebung eigenständig interagieren sollen. Die Detektion interessanter Objekte ist thematisch mit dem Objektlokalisierungsproblem (Grauman und Leibe [2], S. 8) verwandt. Bei diesem geht es darum, in einem Bild alle Objekte einer bestimmten Objektkategorie ausfindig zu machen. Die Art der Objekte, die es zu detektieren gilt, hängt dabei von der konkreten Aufgabenstellung ab. Denkt man beispielsweise an eine Szene aus dem Straßenverkehr, kann es sich bei solchen Objekten um Autos, Fahrräder, Fußgänger etc. handeln. Bei dieser Formulierung der Problemstellung gilt es also, möglichst *alle* Objekte einer Kategorie zu lokalisieren, während es bei der hiesigen Betrachtung darum geht, die *interessanten* Objekte zu finden. Dabei ist die These von zentraler Bedeutung, dass die interessanten Objekte einer Szene visuell salient sind [3]. Visuelle Salienz ist die Eigenschaft bestimmter Bildinhalte, sich aufgrund ihrer visuellen Eigenschaften von ihrem Kontext abzuheben. Die These besagt also, dass interessante Objekte solche Objekte sind, die aufgrund ihrer visuellen Eigenschaften eine exponierte Stellung in der Szene einnehmen. In dieser Arbeit wird diese These in Hinblick auf eine „aufmerksamkeitsbasierte Objektlokalisierung“ untersucht.

1.1 EINORDNUNG DES THEMAS

Bezüglich der Themenzuordnung ist zu beachten, dass die Begriffe *Computer-Vision*, *Machine-Vision* und *Bildverarbeitung* in der Literatur gelegentlich synonym verwendet werden. Bildverarbeitung ist der Oberbegriff für mathematische und signaltheoretische Methoden zur Manipulation und Repräsentation von Bildern. Der Begriff Computer-Vision bezeichnet hingegen ein Forschungsgebiet, das Techniken der Bildverarbeitung mit Methoden des *maschinellen Lernens* und der *künstlichen Intelligenz* mit dem Ziel vereinigt, Bildsemantiken automatisiert zu entschlüsseln. Einfacher ausgedrückt geht es um das Verstehen von Bildinhalten. Die Bildverarbeitung stellt hier folglich nur einen Teilaspekt einer größeren Betrachtung dar. Hinzu kommen weitere Teilbereiche, die als eigenständige Themengebiete aufzufassen sind. Des Weiteren gilt es, die Begriffe Machine-Vision und Computer-Vision voneinander abzugrenzen. Die Unterscheidung ist nicht offensichtlich, da es bei beiden Themengebieten um die Auswertung von Bildinhalten geht. Bei Vernon [4] werden diesbezüglich die unterschiedlichen Zielsetzungen hervorgehoben. Demnach steht bei Machine-Vision der Einsatz optischer Sensoren bei industriellen Fertigungsprozessen im Vordergrund. Hieraus Ergeben sich Schwerpunkte wie beispielsweise optisch gestützte Qualitätskontrolle und Prozessautomatisierung. Demgegenüber steht bei Computer-Vision die Funktionsweise des menschlichen Sehsystems mit der Fragestellung im Vordergrund, wie eine vergleichbare Leistungsfähigkeit in technischen Systemen erreicht werden kann. Dementsprechend stehen Themen wie künstliche Intelligenz und

maschinelles Lernen stark im Vordergrund und es fließen auch Erkenntnisse aus biologischer und psychologischer Forschung in die technischen Modelle mit ein.

Das in dieser Arbeit zentrale Thema der visuellen Aufmerksamkeit im Zusammenhang mit der Objektlokalisierung ist entsprechend der obigen Ausführungen dem Bereich Computer-Vision zuzuordnen, da es um eine technische Anwendung geht, die sich an der menschlichen Wahrnehmung orientiert.

1.2 MOTIVATION UND ZIELE

Wie im letzten Abschnitt erläutert wurde, lässt sich diese Arbeit dem übergeordneten Themenbereich der Computer-Vision zuordnen. Eine wichtige Motivation für die Forschung auf diesem Gebiet sind die Anwendungsmöglichkeiten in der Robotik, mit der Zielsetzung, intelligente Systeme zu entwickeln, die mit ihrer Umgebung eigenständig interagieren können. Hierzu ist der Einsatz einer geeigneten Sensorik erforderlich. Kamerasensoren stellen eine vergleichsweise günstige Möglichkeit dar und liefern umfangreiche Umgebungsinformationen. Für eine Auswertung dieser Informationen können dann Methoden der Computer-Vision eingesetzt werden. Darüber hinaus gibt es in der Praxis Anwendungsmöglichkeiten, die sich durch integrierte Kamerasensoren in vielen Mobilgeräten und eingebetteten Systemen ergeben. Ein bekanntes Beispiel hierfür ist die Gesichtsdetektion (Grauman und Leibe [2], S. 104 ff.), die in unterschiedlichen Systemen eingesetzt werden kann. Ein Beispiel hierfür ist die personalisierte Benutzerinteraktion beispielsweise im Bereich der Heimautomation. Da die Leistungsfähigkeit von Mobilgeräten und eingebetteten Systemen auch in Zukunft weiter steigen wird, dürfte die Bedeutung solcher Anwendungen in Zukunft weiter zunehmen.

Für verschiedene Problemstellungen aus dem Bereich Computer-Vision sind effiziente Verfahren bekannt, die für praktische Anwendungen eine hinreichende Qualität bieten. Diese praxistauglichen Verfahren haben aber einen eher begrenzten Fokus, d.h. sie sind auf eine stark eingegrenzte Problemstellung spezialisiert, wie beispielsweise das detektieren einer ganz bestimmten Objektkategorie. Im Vergleich zu solchen Verfahren ist das menschliche Sehsystem wesentlich leistungsfähiger. So ist der Mensch in der Lage, komplexe Bildinhalte in den unterschiedlichsten Situationen mühelos zu entschlüsseln. Er kann eine Vielzahl verschiedener Objekte unter den unterschiedlichsten Rahmenbedingungen bezüglich Lage, Entfernung und Lichtverhältnisse erkennen. Der Grund für den begrenzten Fokus praktischer Anwendungen liegt zum einen in der hohen Komplexität und aufwendigen Entwicklung solcher Verfahren. Zum anderen skalieren viele der eingesetzten Methoden bei steigenden Anforderungen schlecht. So steigt im Allgemeinen die Komplexität und Fehleranfälligkeit von Objektdetektoren mit der Anzahl der Objektkategorien, zwischen denen unterschieden werden soll. Gleiches gilt für den Entwicklungsaufwand. Solche Verfahren effizienter zu gestalten, stellt ein lohnenswertes Ziel dar. Ein Schlüssel dazu ist das eingangs

erwähnte Prinzip der visuellen Aufmerksamkeit. Für dieses gibt es in der Praxis verschiedene Anwendungsmöglichkeiten, wie beispielsweise intelligente Kameras [5], Bild- und Videokompression [6], Objekt-Tracking [7], Objektsuche [8] [9] und Objektdetektion (siehe Kapitel 5). In dieser Arbeit wird die Betrachtung auf das Problem der Objektdetektion gelegt. Eine wichtige Motivation stellt dabei die Arbeit von Elazary und Itti [3] dar. Die Autoren stellen die eingangs schon erwähnte These auf, dass interessante Objekte visuell salient sind. Der Begriff *interessant* lässt zunächst einmal reichlich Interpretationsspielraum zu. Es lässt sich nicht scharf definieren, welche Kriterien Objekte erfüllen müssen, um interessant zu sein. Die Autoren lösen dass Problem, indem sie eine öffentliche Bilddatenbank (*LabelMe* [10]) verwenden, bei der unabhängige Bearbeiter ohne bestimmte Vorgaben Objekte gelabelt haben. Die Annahme ist nun, dass die Bearbeiter mangels konkreter Vorgaben solche Objekte bearbeitet haben, die sie als interessant erachtet haben. Auf den gelabelten Bildern wurde nun ein Verfahren angewendet, das saliente Punkte in Bildern detektiert. Bei den Experimenten stellte sich heraus, dass ein Großteil der salienten Punkte tatsächlich auf gelabelten Objekten liegt. Diese Beobachtung führt zu der These, dass der Einsatz von Salienzmethoden auf die Problemstellung der Objektdetektion sinnvoll sein kann.

Aufbauend auf dieser These ist es im praktischen Teil dieser Arbeit das Ziel, ein Verfahren zur aufmerksamkeitsbasierten Objektlokalisierung zu entwerfen. Dabei ist dieses Verfahren nicht vornehmlich als Lösungsansatz zum klassischen Lokalisierungsproblem zu verstehen (siehe Abschnitt 3.1). Vielmehr wird eine alternative Formulierung des Problems favorisiert, bei der es darauf ankommt, die interessanten Objekte einer Szene zu lokalisieren. Vorbild hierbei ist das menschliche Sehsystem. Betritt ein menschlicher Betrachter eine Szene, so wird er nicht instantan sämtliche Objekte der Szene identifizieren, sondern er wird aus Gründen der Effizienz auf die wesentlichen Inhalte aufmerksam. Im Gegensatz dazu arbeiten klassische Lokalisierungsverfahren in der Regel auf Basis einer erschöpfende Suche (Grauman und Leibe [2], S. 78 ff.). Bei einer erschöpfenden Suche wird im einfachsten Fall eine Klassifizierung sukzessiv an allen möglichen Positionen und Skalierungen eines größeren Bildes durchgeführt. Dies ist zum einen sehr zeitaufwändig. Zum anderen birgt es auch die Gefahr einer erhöhten Anzahl von Fehldetektionen. Dies trifft insbesondere dann zu, wenn eine große Anzahl unterschiedlicher Objektkategorien betrachtet wird. Es gibt in der Literatur verschiedene Ansätze, die erschöpfende Suche effizienter zu gestalten. Eine entsprechende Betrachtung erfolgt in Kapitel 3. Der aufmerksamkeitsbasierte Ansatz, der in dieser Arbeit verfolgt wird, ist in Abbildung 1.2.1 illustriert. Es soll im Rahmen dieser Arbeit ein auf dem Aufmerksamkeitsprinzip basierendes Verfahren zur Objektlokalisierung entstehen, das sich an der visuellen Wahrnehmung des Menschen orientiert. Dieses setzt sich aus einer unbewussten und zeiteffizienten Phase der Aufmerksamkeitserzeugung und einer bewussten Verarbeitungsphase

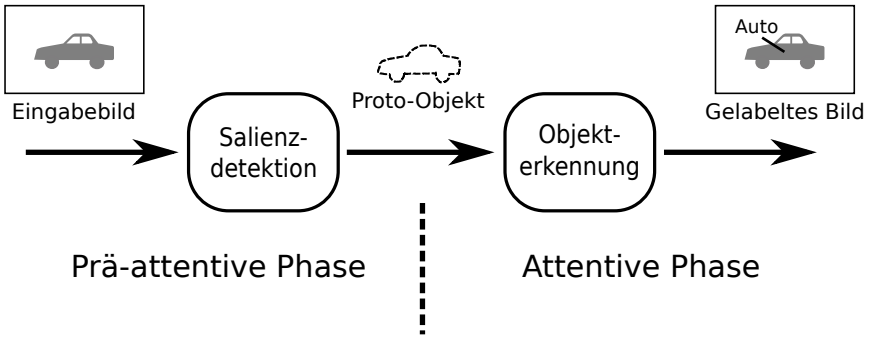

Abbildung 1.2.1: Der Objektdetektor als zweistufiger Prozess. Im ersten Schritt ermittelt ein Salienzdetektor hervorstechende Bildinhalte, aus denen Proto-Objekte erzeugt werden. Im zweiten schritt wird die Semantik der Proto-Objekte mit mittels eines Objektklassifizierungsverfahrens interpretiert.

zusammen. Im ersten Schritt soll ein Salienzdetektor eingesetzt werden, dessen Aufgabe es ist, aufbauend auf einem Aufmerksamkeitsmodell saliente Bereiche des Eingabebildes zu detektieren. Dieser filtert aus dem Eingabebild hervorstechende Bereiche heraus, deren visuellen Eigenschaften interessante Inhalte versprechen. Ein solcher Bereich wird als Proto-Objekt (Jenkin und Harris [1], S. 172 f.), also ein Objekt mit noch unbekannter Bedeutung, bezeichnet. Diese Proto-Objekte werden dann in einem zweiten Schritt klassifiziert und es wird ihnen ein Label zugeordnet. Dieser Schritt entspricht der bewussten Wahrnehmung des Menschen. Analog zur menschlichen Wahrnehmung wird der Klassifizierungsprozess als der komplexe, im Bewusstsein stattfindende Prozess aufgefasst. Der Suchraum wird bei diesem Ansatz analog zum Prinzip der selektiven Aufmerksamkeit eingeschränkt und die Klassifizierung auf die interessanten Teile des Bildes konzentriert. Bei der Klassifizierung werden insbesondere zwei Ziele verfolgt. Zum einen soll der Klassifizierer generisch einsetzbar sein. *Generisch* bedeutet in diesem Zusammenhang, dass das Verfahren, das dem Klassifizierer zugrunde liegt, nicht auf bestimmte Objektklassen zugeschnitten ist, sondern sich prinzipiell auf unterschiedlichste Objektklassen anwenden lässt. Ein weiteres Ziel ist eine gute Skalierbarkeit bezüglich der Anzahl der Objektkategorien, d.h. es soll eine größere Anzahl an Objektkategorien gleichzeitig detektiert werden können. Hierbei spielt auch der Entwicklungsaufwand eine Rolle. Das bedeutet, dass der zusätzliche Arbeitsaufwand möglichst gering ausfallen soll, wenn dem System eine neue Objektkategorie zur Erkennung hinzugefügt werden soll. Das Lokalisierungsverfahren soll schließlich an einem schwierigen Benchmark mit komplexen Szenen getestet werden, um den praktischen Nutzen für den Einsatz von Salienzmethoden bei der Objektlokalisierung zu untersuchen und die entsprechenden Vor- und gegebenenfalls auch Nachteile herauszuarbeiten.

1.3 AUFBAU DER ARBEIT

Diese Arbeit ist in insgesamt acht Kapitel unterteilt. Im Anschluss an diese Einleitung folgt in Kapitel 2 die Erläuterung wichtiger Grundlagen, die zum Verständnis der späteren Themen erforderlich sind. Betrachtet werden Methoden der Bildverarbeitung, sowie Themen aus dem Bereich der Mustererkennung und des maschinellen Lernens. In Kapitel 3 geht es um das Thema der Objektdetektion. Neben einigen wichtigen Grundlagen werden die Vor- und Nachteile verschiedener Lokalisierungsstrategien im Hinblick auf den aufmerksamkeitsbasierten Ansatz betrachtet. Kapitel 4 behandelt das Thema der visuellen Aufmerksamkeit. Ausgang des Kapitels bildet die Betrachtung des menschlichen Sehsystems. Es werden wichtige Erkenntnisse aus den entsprechenden Forschungsgebieten aufgezeigt, die die Grundlage für die Ableitung technischer Aufmerksamkeitsmodelle darstellen. In Kapitel 5 werden verwandte Arbeiten aus der Literatur betrachtet. Die Betrachtung richtet sich zum einen auf verschiedene mathematische Aufmerksamkeitsmodelle. Dabei wird der Schwerpunkt insbesondere auf die Detektion salienter Regionen gelegt. Zum anderen werden Objektlokalisierungsverfahren betrachtet, bei denen auf die eine oder andere Weise Prinzipien der selektiven Aufmerksamkeit bzw. Methoden der Salienzdetektion eingeflossen sind. Kapitel 6 beschreibt die Umsetzung des praktischen Teils dieser Arbeit. Alle verwendeten Methoden zur aufmerksamkeitsbasierten Objektlokalisierung werden hier ausführlich erläutert und begründet. Kapitel 7 behandelt schließlich die Evaluierung des Lokalisierungsverfahrens. Es werden die durchgeführten Experimente und die dabei verwendeten Evaluierungsmethoden und erzielten Ergebnisse erläutert. Die Arbeit schließt in Kapitel 8 mit einer Zusammenfassung sowie einem Ausblick.

2 GRUNDLAGEN

Zum weiteren Verständnis dieser Arbeit sind einige Vorkenntnisse erforderlich, die in diesem Kapitel aufgeführt und in ihren Grundzügen erläutert werden. Darüber hinaus werden zu den einzelnen Themen Literaturhinweise angegeben, denen für eine weitere Vertiefung des jeweiligen Themas gefolgt werden kann. Bei allen der hier behandelten Themen werden Grundkenntnisse der digitalen Signal- und Bildverarbeitung vorausgesetzt, wie sie in zahlreichen Lehr- und Nachschlagewerken vermittelt werden (bspw. Gonzalez und Woods [11]). Für ein tieferes Verständnis sind darüber hinaus Vorkenntnisse auf dem Gebiet der Mustererkennung und des maschinellen Lernens hilfreich (bspw. Bishop [12]). Dieses Kapitel ist in die Abschnitte Scale-Space-Theorie (Abschnitt 2.1), Bildsegmentierung (Abschnitt 2.2) und Mustererkennung (Abschnitt 2.3) unterteilt. Die zum Thema Mustererkennung behandelten Themen stellen die Grundlage zum Thema Objektdetektion dar, das anschließend in Kapitel 3 betrachtet wird. Nicht betrachtet werden hier die Grundlagen zur visuellen Aufmerksamkeit, da diesem Thema ein eigenes Kapitel gewidmet ist (Kapitel 4).

2.1 SCALE-SPACE-THEORIE

In einer unkontrollierten Umgebung können die Entfernungen von Objekten variieren. Einige Merkmale eines Objekts lassen sich besser aus der Nähe bei einer entsprechend hohen Bildauflösung, andere Merkmale besser aus einer größeren Entfernung zu einem Objekt betrachten. Es ist daher wichtig, unterschiedliche Skalierungsstufen modellieren zu können. Ein entsprechendes Modell wird mit der Scale-Space-Theorie nach Lindeberg [13] aufgestellt. Wie später noch zu sehen sein wird, wird dieses Modell in dieser Arbeit in mehreren Bereichen bezüglich Salienzdetektion, Merkmalsextraktion und Bildsegmentierung eingesetzt. Das Modell basiert auf einer Darstellung mittels einer sogenannten *Gauß-Pyramide*, die im Abschnitt 2.1.1 beschrieben wird. Hierauf aufbauend kann die Merkmalsbetrachtung auf unterschiedlichen Skalierungsstufen erfolgen, wie in Abschnitt 2.1.2 am Beispiel der *Difference-of-Gaussian*-Merkmale verdeutlicht wird.

2.1.1 *Gauß-Pyramide*

Bei dem hier betrachteten Modell wird der Scale-Space mittels einer Gauß-Pyramide (Gonzalez und Woods [11], S. 351) aufgespannt. Jede Stufe der Pyramide repräsentiert dabei eine Skalierungsstufe. Das Ausgangsbild in seiner vollen

Auflösung stellt die erste Stufe der Pyramide dar. Die weiteren Stufen werden erstellt, indem auf die vorangehende Stufen jeweils ein Tiefpassfilter (TP) zur Glättung angewendet und anschließende eine Unterabtastung durchgeführt wird. Bei der Unterabtastung wird in der Regel ein Faktor von Zwei ($\downarrow 2$) angesetzt, da sich dies am effizientesten realisieren lässt. Zur Tiefpassfilterung wird ein Gauß-Filter eingesetzt. Anhand informationstheoretischer Erwägungen hat Lindeberg formal dargelegt, dass es sich beim Gauß-Filter um das ideale Filter für diesen Zweck handelt (siehe [13]). Die Gauß-Pyramide wird entsprechend der folgenden Gleichung aufgestellt:

$$G_l(x,y) = \begin{cases} I(x,y), \text{wenn } l = 1 \\ \sum_{m=-\lfloor \frac{k}{2} \rfloor}^{\lfloor \frac{k}{2} \rfloor} \sum_{n=-\lfloor \frac{k}{2} \rfloor}^{\lfloor \frac{k}{2} \rfloor} w(m,n) G_{l-1}(2x+m, 2y+n), \text{sonst} \end{cases} \tag{2.1.1}$$

Hier stellt l die Skalierungsstufe, w die Filtermaske und k ihre Größe dar. Da Gauß-Filter symmetrisch und separierbar sind, lässt sich der Prozess durch zwei eindimensionale Faltungen effizient umsetzen. Hierbei ist die Approximation eines Gauß-Filters beispielsweise in der Form $\frac{1}{16}[1\,4\,6\,4\,1]$ für praktische Zwecke in der Regel ausreichend. Abbildung 2.1.1 illustriert das Vorgehen anhand eines Beispiels. Hier wird verdeutlicht, wie durch das Herausfiltern der feinen, hochfrequenten Details und anschließender Reduktion der Auflösung sukzessiv aus dem Bild herausgezoomt wird.

2.1.2 *Difference-of-Gaussian*

Mit *Difference-of-Gaussian* (DoG) [13] wird die Differenz zweier unterschiedlich stark geglätteter Versionen eines Bildes entsprechend der folgenden Formel bezeichnet.

$$DoG_{\sigma_1,\sigma_2}(x,y) = G(x,y,\sigma_1) - G(x,y,\sigma_2) \tag{2.1.2}$$

Mit G wird hier die Gauß-Funktion bezeichnet. Bei den Gauß-Filtern wird die Stärke der Glättung jeweils durch die Standardabweichung, σ_1 bzw. σ_2, bestimmt. Mathematisch wird hierdurch ein Bandpassfilter realisiert, bei dem das Frequenzband durch das Verhältnis von σ_1 zu σ_2 bestimmt wird. Auf diese Weise lassen sich gezielt die Details für einen bestimmten Ortsfrequenzbereiche isolieren. Wird dieses Verfahren auf eine Gauß-Pyramide erweitert, lässt sich eine umfassende Merkmalsbetrachtung für unterschiedliche Bildkoordinaten, Skalierungsstufen und Ortsfrequenzbereiche durchführen. In [14] wird dieses Vorgehen zur Realisierung eines Keypoint-Detektors (Grauman und Leibe [2], S. 13 ff.) eingesetzt. Die Keypoints entsprechen dabei lokalen Maxima der Merkmalswerte innerhalb des Scale-Space. In [15] werden DoG-Filter bei der Salienzdetektion eingesetzt (siehe Abschnitt 4.3).

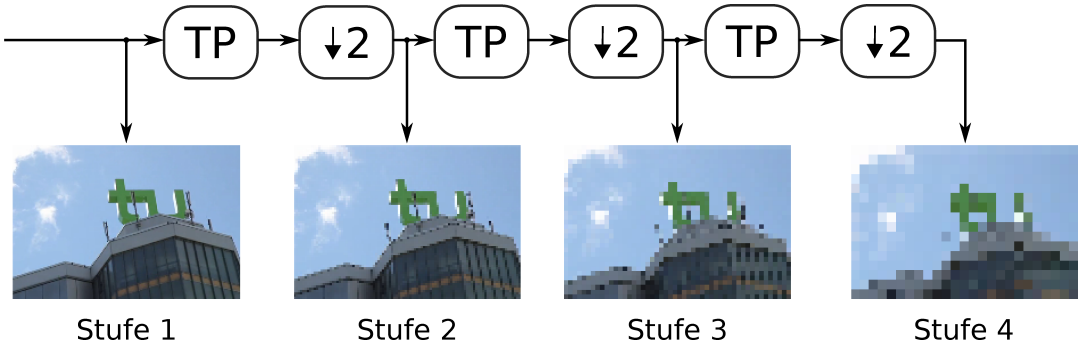

Abbildung 2.1.1: Gauß-Pyramide: Beispiel mit vier Skalierungsstufen.

2.2 BILDSEGMENTIERUNG

Bei der Bildsegmentierung (Gonzalez und Woods [11], S. 567) ist es das Ziel, die logisch zusammenhängenden Bereiche eines Bildes als Segmente auszuweisen. Es handelt sich hierbei um eine Form der Komplexitätsreduktion, bei der eine verhältnismäßig große Zahl an Pixeln in eine vergleichsweise geringe Zahl an aussagekräftigen Segmenten überführt wird. Im Allgemeinen handelt es sich bei Segmentierungsproblemen um schlecht gestellte Probleme. In vielen Fällen gibt es mehrere Möglichkeiten, Pixelgruppen sinnvoll zusammenzufassen. Ist dies aber der Fall, so lässt sich weder eine eindeutige Lösung finden, noch lässt sich eine gefundene Lösung nach allgemeingültigen Kriterien auf ihre Richtigkeit überprüfen. Ja nach Problemstellung ist jedoch eine Überprüfung mittels Expertenwissens möglich. Typischerweise wird hierbei eine gewisse Anzahl von Bildern nach bestimmten Vorgaben von Personen händisch segmentiert. Diese Bilder lassen sich dann als Benchmark zur Bewertung eines Segmentierungsverfahrens heranziehen.

Da auf dem Gebiet der Bildsegmentierung seit den Anfängen der digitalen Bildverarbeitung geforscht wird, sind bis heute entsprechend viele Verfahren vorgeschlagen worden. Eine Übersicht hierzu findet sich beispielsweise in [16]. Grundsätzlich lässt sich zwischen generischen Segmentierungsverfahren und solchen, bei denen problemspezifisches Modellwissen eingesetzt wird (Grauman und Leibe [2], S. 112 ff.), unterscheiden. Hierbei gilt, um so konkreter eine Problemstellungen gefasst ist, desto mehr Modellwissen kann auch in ein entsprechendes Segmentierungsverfahren einfließen. Im Kontext der selektiven Aufmerksamkeit sind gerade solche Verfahren von Interesse, bei denen die Einteilung in Segmente ausschließlich anhand grundlegender visueller Merkmale vorgenommen wird. In dieser Arbeit werden deshalb ausschließlich generische Segmentierungsverfahren betrachtet. Zu den typischen Strategien solcher Verfahren gehört es, die Bildpunkte zusammenzufassen, die eine gleichartige Fläche beschreiben (Gonzalez und Woods [11], S. 612 ff.), oder sie anhand ihrer Kantenübergänge von benachbarten Flächen abzugrenzen (Gonzalez und Woods [11], S. 585 ff.). Auch eine

Kombination von Regionen- und Kanteninformationen ist möglich. Eine Übersicht zu solchen kombinierten Ansätzen findet sich in [17]. In der Praxis können sich bei der Segmentierung Probleme ergeben, wenn Eigenschaften bezüglich Kontrast, Helligkeit, Detailreichtum etc. variieren. Die Qualität der Ergebnisse kann entsprechend von Bild zu Bild unterschiedlich ausfallen. Man spricht in diesem Zusammenhang von einer Übersegmentierung, wenn zu viele Segmente erzeugt werden, beziehungsweise von einer Untersegmentierung bei zu wenigen Segmenten.

Eine weitere Unterscheidung lässt sich zwischen solchen Verfahren treffen, die die gesamte Bildfläche in Segmente aufteilen (Gonzalez und Woods [11], S. 587 ff.), und solchen Verfahren, die ausgehend von einem Startpunkt lediglich einen einzelnen Bereich heraussegmentieren (Gonzalez und Woods [11], S. 585 ff., S. 612 ff.). Die letztere Strategie kann verwendet werden, um das zu einem salienten Punkt gehörende Objekt zu segmentieren. Ein entsprechendes Verfahren kann beispielsweise auf dem Region-Growing-Prinzip (Gonzalez und Woods [11], S. 613 ff.) basieren, bei dem das Segment ausgehend von einem Startpunkt sukzessiv durch die Hinzunahme benachbarter, gleichartiger Pixel anwächst. Das Aufteilen der gesamten Bildfläche in Segmente ist hingegen für eine andere Strategie der Salienzdetektion von Bedeutung, bei der zunächst die Segmente ermittelt und anschließend deren Salienz bewertet wird (eine ausführlich Erläuterung hierzu folgt in Kapitel 4). Zu dieser Kategorie gehören graphenbasierte Verfahren, wie sie in Abschnitt 2.2.1 betrachtet werden. In Abschnitt 2.2.2 wird das Prinzip der multiskalaren Segmentierung erläutert, welches das Segmentierungsproblem auf die Betrachtung mehrerer Skalierungsstufen erweitert.

2.2.1 *Graphenbasierte Verfahren*

Bei graphenbasierten Verfahren (Gonzalez und Woods [11], S. 591 ff.) wird das Bild als Graph modelliert. Einzelne Pixel stellen die Knoten dar und die Übergänge zwischen benachbarten Pixeln werden durch Kanten repräsentiert. Die Segmentierung lässt sich so als Partitionierungsproblem formulieren, bei dem es das Ziel ist, den Graphen unter Anwendung bestimmter Kriterien so zu partitionieren, dass die logisch zusammenhängenden Bereiche jeweils durch einen Teilgraphen repräsentiert werden. Eine Übersicht zu graphenbasierten Verfahren findet sich in [18]. Felzenszwalb und Huttenlocher stellen in [19] ein graphenbasiertes Segmentierungsverfahren vor, bei dem Segmente durch den Aufbau minimaler Spannbäume gebildet werden. Jeder Pixel wird durch einen Knoten repräsentiert. Kanten bestehen zwischen benachbarten Pixeln jeweils in horizontaler, vertikaler und diagonaler Richtung. Dabei bestimmen sich die Kantengewichte zu $d(I(j,i), I(j,i+1))$, $d(I(j,i), I(j+1,i))$ bzw. $d(I(j,i), I(j+1,i+1))$. Die Funktion $d(p,q)$ gibt die Differenz zwischen zwei Bildpunkten,

$p = (p_r, p_g, p_b)$ und $q = (q_r, q_g, q_b)$, an. Sie ist ein Maß dafür, wie stark sich zwei Pixel voneinander unterscheiden und bestimmt sich zu

$$d(p,q) = \sqrt{(p_r - q_r)^2 + (p_g - q_g)^2 + (p_b - q_b)^2}. \tag{2.2.1}$$

Hierbei stehen r, g und b für den Rot- Grün- bzw. Blauanteil. Initial stellt jeder Knoten für sich einen minimale Spannbaum dar, der eine Region aufspannt, die aus einem Pixel besteht. Die Kanten werden nun anhand ihres Gewichts sortiert und anschließend in nicht-absteigender Reihenfolge verarbeitet. Sofern die jeweils betrachtete Kante zwischen zwei Spannbäumen liegt, also zwei Regionen voneinander trennt, wird entschieden, ob die beiden beteiligten Regionen vereinigt werden, oder ob die Kante als echte Grenze eingestuft wird. Die Vereinigung geschieht, indem die Spannbäume über die betrachtete Kante zu einem neuen Spannbaum verbunden werden. Da dieses Vorgehen dem Algorithmus von Kruskal [20] entspricht, ist der neue Spannbaum wiederum minimal. Um zu entscheiden, ob zwei Regionen vereinigt werden, wird die innere Differenz der Regionen mit dem Gewicht der betrachteten Kante verglichen. Die innere Differenz ist definiert als das maximale Kantengewicht, $G_{S,max}$, des minimalen Spannbaums eines Segments, S. Der Schwellwert, T_S, bestimmt sich dann zu

$$T_S = G_{S,max} + \frac{c}{|S|}. \tag{2.2.2}$$

Dabei stellt c einen frei wählbaren Parameter und $|S|$ die Größe des Segments S dar. Ist das Gewicht der betrachteten Kante kleiner als die beiden Schwellwerte der beteiligten Regionen, werden diese vereinigt. Zur effizienten Umsetzung des Verfahrens müssen nicht die Spannbäume insgesamt, sondern nur das jeweils höchste Gewicht der Spannbäume zwischengespeichert werden. Die Komplexität der Verfahrens, $O(|E| \log(|E|))$, wird durch das Sortieren der Kanten bestimmt, wobei $|E|$ der Anzahl der Kanten entspricht.

2.2.2 *Multiskalare Segmentierung*

Das Konzept der multiskalaren Segmentierung beruht, wie der Name vermuten lässt, auf der Durchführung von Segmentierungsmethoden auf unterschiedlichen Skalierungsstufen. Zur Modellierung der Skalierungsstufen kann das in Abschnitt 2.1 beschriebene Scale-Space-Modell eingesetzt werden. Beispiele für den Einsatz von Scale-Space-Segmentierung finden sich unter anderem bei Themen wie der bildgebenden Diagnostik in der Medizin [21] oder bei der Auswertung von Luftbildern [22]. Für das Segmentieren im Rahmen der Objektdetektion ist die Betrachtung mehrerer Skalierungsstufen interessant, da sie es erlaubt, flexibler mit dem Problem variierender Objektgrößen bzw. Entfernungen umzugehen. Kleine Objekte können besser auf einer feinen Detailstufe segmentiert werden, große hingegen auf einer groben Stufe. Das Konzept ist an kein bestimmtes

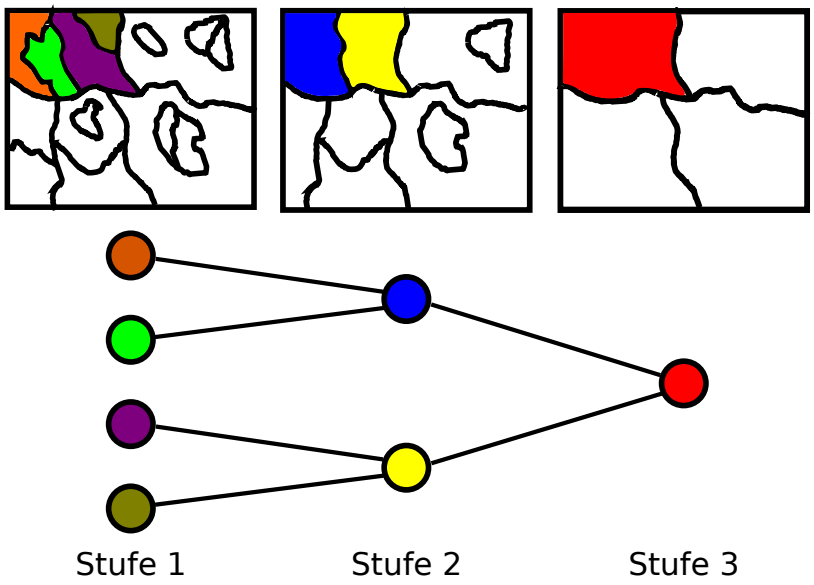

Abbildung 2.2.1: Prinzipskizze der hierarchischen, multiskalaren Segmentierung.

Segmentierungsverfahren gebunden, sondern als allgemeines Prinzip zu verstehen. Im einfachsten Fall kann das Verfahren wiederholt auf den unterschiedlichen Skalierungsstufen angewendet werden. Das Vorgehen eignet sich, um das Problem der Übersegmentierung zu adressieren. Durch das Herauszoomen können Details reduziert werden, was bei der Segmentierung zu weniger Segmenten führt. Darüber hinaus werden bei einigen multiskalaren Segmentierungsverfahren auch die hierarchischen Beziehungen zwischen den Segmenten herausgestellt. Dabei setzt sich ein Segment einer gröberen Skalierungsstufe aus mehreren Segmenten einer feineren Stufe zusammen. Diese Abhängigkeiten können mit Hilfe von Graphen modelliert werden, wie es in Abbildung 2.2.1 illustriert wird.

2.3 MUSTERERKENNUNG

Unter einem Muster (Bishop [12], S. 1 ff.) ist ganz allgemein eine Regelmäßigkeit in einer Datenmenge zu verstehen. Solche Regelmäßigkeiten können genutzt werden, um bei unbekannten Dateneingaben Rückschlüsse auf deren Inhalt zu ziehen. Das bedeutet, dass durch das Auftreten einer Regelmäßigkeit gefolgert werden kann, dass ein bestimmter Inhalt vorliegen muss. Häufig ist hierbei die Zielsetzung, spezifische Inhalte wiederzuerkennen und zu benennen. Dies wird als Mustererkennung (Bishop [12], S. 1 ff.) bezeichnet. Ein Muster muss charakteristische Eigenschaften besitzen, die eine gewisse Alleinstellung gewährleisten und somit eine Wiedererkennung ermöglichen. Solche Eigenschaften werden als Merkmale (Bishop [12], S. 2) bezeichnet. Ein Muster kann durch eine Menge an Merkmalen und deren örtlichen bzw. zeitlichen Beziehungen zueinander beschrieben werden. Ein Beispiel für den Einsatz von Mustererkennung ist die

Spracherkennung (Bishop [12], S. 605 ff.). In diesem Zusammenhang kann das betrachtete Ereignis ein bestimmtes gesprochenes Wort in einer Audiosequenz sein. Das Muster spiegelt sich dann in bestimmten Merkmalen wieder, die Eigenschaften bezüglich der Klangformung des Wortes beschreiben. Hierbei ist es wichtig, mögliche Variationen zu berücksichtigen. Beispielsweise variiert der Klang eines Wortes, je nachdem wie schnell, wie laut und von welcher Person es gesprochen wird. Darüber hinaus können störende Einflüsse, wie Hintergrundgeräusche oder eine wechselnde Aufnahmequalität, hinzukommen. Es gilt also, die wiederkehrenden Gemeinsamkeiten unter Berücksichtigung der möglichen Variationen herauszustellen.

Die Art der Daten, die bei der Mustererkennung betrachteten werden, kann je nach Problemstellung recht unterschiedlich sein. Je nach Datenformat kann die Betrachtung auf örtlich oder zeitlich auftretenden Ereignissen liegen. Die unterschiedlichen Problemstellungen haben aber in aller Regel gemeinsam, dass eine hochdimensionale, verhältnismäßig komplexe und unübersichtliche Dateneingabe in eine niedrigdimensionale, aussagekräftige und wohldefinierte Ausgabe überführt werden soll. Bezüglich der Art der Ein- und Ausgabe können Mustererkennungsprobleme in unterschiedliche Kategorien eingeteilt werden. Zu den wichtigsten zählen Klassifizierungs- und Regressionsprobleme. Bei Klassifizierungsproblemen (Bishop [12], S. 179 ff.) stellt die Eingabe eine Beobachtung dar, die einer von mehreren möglichen Klassen zugeordnet werden soll. Bei Regressionsproblemen (Bishop [12], S. 137 ff.) hingegen soll der Grad bestimmter Eigenschaften der Beobachtung durch kontinuierliche Ausgabewerte beschrieben werden. Im Bereich Computer-Vision werden bei der Mustererkennung Daten visueller Natur betrachtet. Diese liegen in der Regel in Form von Bildern oder Videosequenzen vor. Konkrete Beispiele sind die Schrifterkennung (Bishop [12], S. 614 f.) oder auch die in dieser Arbeit betrachtete Objekterkennung (Grauman und Leibe [2], S. 1 ff.). Die Objekterkennung kann als Klassifizierungsproblem aufgefasst werden, bei der ein Eingabebild einer von mehreren Objektkategorien zugeordnet werden soll (eine genaue Betrachtung der Problemstellung erfolgt in Kapitel 3). Hierbei wird vorausgesetzt, dass sich die Objekte einer Kategorie bestimmte visuelle Eigenschaften teilen, also ein gemeinsames Muster aufweisen. In Abbildung 2.3.1 wird ein typisches Schema zur Lösung eines Klassifizierungsproblems dargestellt. Der erste Schritt ist die Normalisierung der Eingabe, die dem Ausgleich verschiedener Rahmenbedingungen dient, die bei der Erfassung der Eingabedaten variieren können. Bei Bildern können solche Maßnahmen beispielsweise eine Vereinheitlichung der Abtastrate, des Wertebereichs, der durchschnittlichen Helligkeit oder des Kontrastes umfassen. Es folgt anschließend die Extraktion der Merkmale. Mögliche Vorgehensweisen hierzu werden in Abschnitt 2.3.1 erörtert. Die Merkmale werden anschließend im Klassifizierungsprozess einer Klasse zugeordnet. Hierzu werden die Merkmale mit bekannten Mustern abgeglichen. Bei der Klassifizierung spielen Methoden des maschinellen

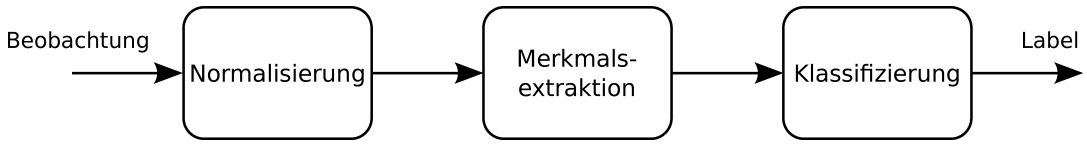

Abbildung 2.3.1: Schema zur Lösung eines Klassifizierungsproblems.

Lernens (Bishop [12], S. 2) eine starke Rolle, bei denen es um die Anpassung von Modellparametern mit Hilfe von Beispieldaten geht. Beide Themen werden in Abschnitt 2.3.2 behandelt.

2.3.1 *Merkmalsextraktion*

Wie eingangs erwähnt, wird mit dem Begriff *Merkmal* eine charakteristische Eigenschaft einer Dateneingabe, wie bspw. einem Bild- oder Audiosignal, bezeichnet. *Merkmalsextraktion* (Bishop [12], S. 2) wird der Prozess genannt, der nach einem bestimmten Verfahren die betrachteten Merkmale aus der Dateneingabe ermittelt und in eine fest definierte, vektorielle Darstellung überführt. Diese wird dann als Merkmalsvektor oder Merkmalsdeskriptor bezeichnet. Verfahren zur Merkmalsextraktion werden im Themenbereich Computer-Vision häufig eingesetzt. In dieser Arbeit ist das Thema neben der hier betrachteten Mustererkennung auch bei der Salienzdetektion im Hinblick auf die Extraktion von Salienzmerkmalen von Bedeutung. Bei der Salienzdetektion werden jedoch andere Anforderungen an die betrachteten Merkmale gestellt (hierzu später mehr in Kapitel 4).

Im Folgenden wird eine Dateneingabe mit I bezeichnet. Ohne Beschränkung der Allgemeinheit wird angenommen, dass es sich bei I um ein n-Tupel reeller Zahlen, $I \in \mathbf{R}^n$, handelt. Für mehrdimensionale Datenstrukturen, wie Bilder und Ähnliches, lassen sich problemlos entsprechende Darstellungen angeben. Für die Merkmalsextraktion wird nun ein m-dimensionaler Merkmalsraum aufgespannt. Alle betrachteten Merkmale werden durch Merkmalsvektoren in diesem Raum, $\vec{v} \in \mathbf{R}^m$, repräsentiert. Des Weiteren gibt eine Funktion $f : \mathbf{R}^n \mapsto \mathbf{R}^m$ eine Vorschrift an, mit der der Merkmalsvektor aus I extrahiert wird. Die Extraktionsvorschrift sollte so gestaltet sein, dass Merkmale, die einen ähnlichen Inhalt repräsentieren, innerhalb des Merkmalsraums einen geringen Abstand haben. Merkmale, die Inhalte unterschiedlicher Natur repräsentieren, sollten hingegen einen großen Abstand aufweisen. In Hinsicht auf Klassifizierungsprobleme bedeutet dies, das viele Merkmale der gleichen Klasse eine Häufung im Merkmalsraum bilden. Die Annahme dieser Vorraussetzung wird Kompaktheitshypothese (Niemann [23], S. 20) genannt. In diesem Zusammenhang müssen der Extraktionsprozess und die daraus resultierenden Merkmale bestimmte Anforderungen erfüllen. Insbesondere sind dies Anforderungen bezüglich *Komplexität*,

Unterscheidbarkeit und *Robustheit*. Diese Anforderungen sollen im Folgenden erläutert werden.

Die durch die Merkmalsextraktion erstellten Merkmalsvektoren sollten eine geringe Komplexität aufweisen. Das bedeutet, es werden möglichst niedrigdimensionale Merkmalsvektoren bei einem möglichst hohem Gehalt relevanter Informationen angestrebt. Hintergrund dieser Forderung ist der Wunsch, die Komplexität der Dateneingabe durch eine Konzentration auf die wesentliche Inhalte zu reduzieren. Eine kompakte Darstellung verspricht einen effizienten Klassifizierungsprozess. Eine Reduktion der Komplexität lässt sich auch in einem eigenen Prozess erreichen, der der Merkmalsextraktion nachgelagert wird. Hierbei wird die Komplexität der Merkmalsvektoren durch die Minimierung von Redundanzen nachträglich reduziert. Ein solcher Prozess wird als *Dimensionalitätsreduktion* (Bishop [12], S. 186 ff.) bezeichnet. Ein bekanntes Verfahren hierzu ist die Hauptkomponentenanalyse (PCA) [24]. Bei diesem Verfahren werden die einzelnen Merkmale als Zufallsvariablen aufgefasst, deren Korrelationen sich durch eine ausreichend große Stichprobe an Merkmalsvektoren näherungsweise bestimmen lässt. Das Ziel besteht darin, die Merkmalsvektoren in eine de-korrelierte Darstellung zu transformieren. Die Reduktion wird anschließend erzielt, indem die Dimensionen mit den geringsten Varianzen vernachlässigt werden.

Mit *Unterscheidbarkeit* (engl. distinctiveness; Grauman und Leibe [2], S. 11) ist die Eigenschaft gemeint, Merkmalsvektoren unterschiedlicher Dateneingaben auseinander halten zu können. Dies ist notwendig, um eine ausreichende Bandbreite an unterschiedlichen Ereignissen ohne die Gefahr der Verwechslung beschreiben zu können. Im Hinblick auf Klassifizierungsprobleme bedeutet dies, dass Merkmalsvektoren, die aus Dateneingaben der gleichen Klasse extrahiert wurden, sich ähneln. Gleichzeitig unterscheiden sie sich aber hinreichend von Merkmalsvektoren, die aus Eingaben anderer Klassen extrahiert wurden. Die Schwierigkeit liegt darin, dass es auch innerhalb einer Klasse zu Variationen bezüglich der Merkmale kommt. Das bedeutet, der Abstand zweier Merkmalsvektoren der gleichen Klasse sollte geringer ausfallen, als bei einem Vergleich zweier Merkmalsvektoren unterschiedlicher Klassen.

Mit *Robustheit* (Grauman und Leibe [2], S. 11 f.) wird gefordert, dass die Merkmalsvektoren gegenüber bestimmten äußeren Einflüssen möglichst invariant sind. Das bedeutet, dass die Lage eines Merkmalsvektors im Merkmalsraum nach Möglichkeit nur vom betrachteten Ereignis selbst abhängen sollte, nicht jedoch durch sonstige Einflussfaktoren beeinflusst wird. Solche Faktoren können dem Aufnahmeprozess oder auch einer unkontrollierten Umgebung geschuldet sein, in der die Datenaufnahme durchgeführt wurde. Bei Klassifizierungsproblemen ist im Allgemeinen damit zu rechnen, dass solche Einflüsse unabhängig von einer bestimmten Klasse bei allen Aufnahmen in unterschiedlichen Ausprägungen auftreten können. Bei Bildern gehören zu den wichtigsten Einflussfaktoren die

geometrischen Rahmenbedingungen bzgl. Perspektive, Skalierung, Translation und Rotation der Bildinhalte. Variiert die Projektion der Inhalte von Bild zu Bild, bedeutet dies, dass sich bestimmte relevante Informationen an unterschiedlichen Positionen der Bildmatrix befinden und sich unterschiedlich darstellen können. Eine Steigerung der Robustheit lässt sich prinzipiell in einem Umfang erreichen, in dem die unterschiedlichen Abbildungen affin zueinander sind, d.h. dort, wo sich die Projektionen ineinander überführen und eindeutig umkehren lassen. Man spricht hierbei entsprechend von *affiner Invarianz* (Grauman und Leibe [2], S. 11 f.). Hierbei unterscheidet man im Speziellen zwischen Translations-, Skalierungs- und Rotationsinvarianz. Letztere kann prinzipiell für Rotationen in der Bildebene erreicht werden. Weitere wichtige Einflussfaktoren sind photometrische Rahmenbedingungen, d.h. variierende Lichtverhältnisse. Des Weiteren sind noch Unterschiede im Aufnahmeprozess bzgl. Kontrast, Rauscheinflüsse und Codierungsartefakte (bspw. Aliasing) zu nennen. Je nach Art können einige dieser Einflüsse bereits im Rahmen eines Normalisierungsprozesses ausgeglichen werden, was in der Praxis in der Regel jedoch nur bis zu einem gewissen Grad gelingt.

Bezüglich der oben aufgeführten Anforderungen gilt, dass diese in Abhängigkeit zueinander stehen. So fordert ein hoher Grad an Unterscheidbarkeit auch entsprechend komplexe Merkmalsvektoren, was wiederum die Möglichkeit einer kompakten Merkmalsdarstellung reduziert. Des Weiteren lässt sich im Allgemeinen die Robustheit gegenüber verschiedenen Einflüssen nicht beliebig steigern, ohne dabei auch die Unterscheidbarkeit zu reduzieren. In der Praxis gilt es also einen sinnvollen Kompromiss zwischen den unterschiedlichen Anforderungen zu erzielen. In den folgenden Abschnitten werden diesbezüglich einige bildbasierte Verfahren zur Merkmalsextraktion aufgezeigt.

2.3.1.1 *Histogramme*

Ein trivialer Ansatz zur Merkmalsextraktion ist die Erstellung eines Merkmalsvektors, bei dem jeder Pixelwert einem Vektoreintrag entspricht. Im einfachsten Fall können hierbei Graustufenwerte verwendet werden. Zunächst fällt auf, dass dieser Ansatz keinen Beitrag zur Komplexitätsreduktion leistet, sodass dieser Ansatz von vornherein nur bei einer geringen Bildauflösung in Frage kommen würde. Ein weiteres Problem ist, dass keine Robustheit gegenüber äußeren Einflüssen gewährleistet ist. So wirken sich Änderungen der Lichtverhältnisse und Schattenwürfe im vollen Umfang auf die Merkmalsrepräsentation aus. Teilweise lassen sich solche Probleme durch einen vorgeschalteten Normalisiserungsprozess adressieren. Schwerwiegender noch wirken sich aber geometrische Änderungen aus. Bei einer translatorischen Verschiebung von nur einem Pixel würden sich alle Werte auf anderen Positionen des Merkmalsvektors wiederfinden, was in

der Regel eine nicht unerhebliche Verschiebung der Lage im Merkmalsraum bedeuten würde.

Histogramme (Bishop [12], S.120 ff.) stellen eine einfache Möglichkeit dar, eine Verbesserung bezüglich dieser Probleme zu erzielen. Hierbei wird der betrachtete Wertebereich in mehrere Teilbereiche unterteilt. Diese Bereiche werden als Kanäle bezeichnet (im englischen auch *bin*). Alle Pixel werden nun anhand ihres Wertes einem Kanal zugeordnet. Für jeden Kanal wird die Anzahl der zugeordneten Pixel ermittelt. Aus diesen Häufigkeiten wird schließlich der Merkmalsvektor erstellt. Die Granularität der Kanäle bestimmt dabei die Dimensionalität des Merkmalsvektors. Durch dieses Vorgehen kann eine starke Komplexitätsreduktion erzielt werden, da die Anzahl an Pixeln in der Regel groß ist im Vergleich zur Anzahl der Kanäle. Ein weiterer Vorteil ist die erhöhte Robustheit gegenüber veränderten geometrischen Rahmenbedingungen. Die Lage eines Pixels ist nämlich unerheblich für den Kanal dem er zugeordnet wird. Insbesondere bei Translation und Rotation in der Bildebene wird so eine hohe Robustheit erzielt. Auch wird eine gewisse Robustheit gegenüber Rauscheinflüssen oder leichten Helligkeitsänderungen erzielt. Fällt ein Änderung durch solche Einflüsse nur leicht aus, wird der größere Teil der Pixel immer noch dem gleichen Kanal zugeordnet werden. Neben Graustufen ist es bei Histogrammen auch möglich, Farben oder Gradienten zu betrachten. Sollen hierbei mehrere Werte berücksichtigt werden, bspw. Färbung und Sättigung, so kann dies durch ein mehrdimensionales Histogramm geschehen.

2.3.1.2 *Zellhistogramme*

Wie im letzten Abschnitt verdeutlicht wurde, liegt die Stärke von Histogrammen in ihrer Robustheit gegenüber affinen Transformationen. Dies geht jedoch auf Kosten der örtlichen Informationen, die durch eine Histogrammdarstellung vollständig verloren gehen. Dies bedeutet einen entsprechenden Verlust an Unterscheidbarkeit. Eine diesbezüglich bessere Alternative stellt deshalb der Einsatz von Zellhistogrammen dar. Bei diesen wird die Bildfläche in mehrere Zellen eingeteilt. Für jede Zelle wird dann ein eigenes Histogramm erstellt. Die Histogramme werden durch Konkatenation zu einem Merkmalsvektor zusammengefasst. Da jedes Histogramm eine dedizierte Position in der Merkmalsrepräsentation einnimmt, bleibt ein Teil der spatialen Informationen erhalten. Gleichzeitig reduziert sich jedoch die affine Invarianz. Es handelt sich also um einen Kompromiss zwischen Unterscheidbarkeit und Robustheit. Zellhistogramme werden beispielsweise von Dalal und Triggs [25] für ihre HoG-Merkmale (Histogram of Oriented Gradients) verwendet. Das Zellhistogramm wird dort über zuvor ermittelte Gradientenwerte erstellt. Dabei werden die Gradientenwerte für jede Zelle unter Miteinbeziehung benachbarter Zellen lokal normalisiert. Durch die Normalisierung wird

eine gesteigerte Invarianz gegenüber Änderungen der lokalen Lichtverhältnisse erreicht.

2.3.1.3 *SIFT-Merkmale*

Lowe [14] beschreibt mit seinem Verfahren, das er *Scale-Invariant Feature Transform* (SIFT) nennt, einen dreistufigen Prozess zur Merkmalsextraktion. Im ersten Schritt wird zunächst auf verschiedenen Skalierungsstufen nach Keypoints gesucht (siehe hierzu Abschnitt 2.1.2). Im zweiten Schritt wird für die Kepoints anhand der umgebenden Gradientenwerte eine Orientierung ermittelt. Das Vorgehen erlaubt es, bestimmte Merkmale unabhängig von ihrer Skalierung und Orientierung zu lokalisieren. Die Extraktion der Merkmale, die dann als SIFT-Merkmale bezeichnet werden, erfolgt im dritten Schritt. Betrachtet wird hierbei eine Fläche von 16×16 Pixeln, die durch die Position, Skalierung und Orientierung des jeweils betrachteten Keypoints festgelegt wird. Nach dem Prinzip eines Zellhistogramms wird die Fläche in sechzehn Kacheln mit einer Größe von 4×4 Pixeln eingeteilt. Für jede der Kacheln wird ein Gradientenhistogramm mit acht Kanälen für acht Orientierungen der Gradienten berechnet. Alle Werte werden anschließend mit einer Gauß-Funktion gewichtet, bei der die Standardabweichung, σ, der halben Kachelbreite entspricht. Die Histogramme werden schließlich zu einem Merkmalsvektor mit $16 \times 8 = 128$ Elementen zusammengefasst. Die Untersuchung unterschiedlicher Merkmalsdeskriptoren in [26] attestiert den SIFT-Merkmalen vergleichsweise gute Eigenschaften hinsichtlich eines Merkmalsabgleichs und eine vergleichsweise hohe Robustheit gegenüber photometrischen und affinen Transformationen.

2.3.2 *Klassifizierung und maschinelles Lernen*

Bei der Klassifizierung (Bishop [12], S. 181 ff.) wird eine Dateneingabe einer von mehreren möglichen Klassen zugeordnet. Entsprechende Verfahren werden im Folgenden Klassifizierer genannt. Bei Klassifizierern handelt es sich im Wesentlichen um komplexe Funktionen, denen ein bestimmtes Datenmodell zugrunde liegt. Sie zeichnen sich häufig durch eine große Anzahl konfigurierbarer Parameter und eine hohe Nichtlinearität aus. Die Zielfunktion für ein bestimmtes Klassifizierungsproblem wird durch die geeignete Konfiguration dieser Parameter erreicht. Hierzu werden Methoden des *maschinellen Lernens* (Bishop [12], S. 2 ff.) eingesetzt. Der Vorteil des maschinellen Lernens besteht in der Möglichkeit, mittels Beispieleingaben zu lernen. Im Falle der visuellen Mustererkennung werden hierfür Bilddatenbanken eingesetzt (siehe Kapitel 7). Eine Bilddatenbank muss dabei Beispiele für die zu lernenden Klassen in repräsentativem Umfang enthalten. Der Klassifizierer lernt hieraus, die für die Klassen relevanten Informationen und kann diese dann zur Laufzeit mit unbekannten Eingaben abgleichen.

Dabei wird häufig ein kontinuierlicher Wert für jede Klasse angegeben, der die Wahrscheinlichkeit für die Zugehörigkeit zu dieser Klasse ausdrückt.

Grundsätzlich kann man zwischen zwei Arten des maschinellen Lernens unterscheiden, dem überwachten und dem unüberwachten Lernen (Bishop [12], S. 3). Beim überwachten Lernen (engl. supervised learning) kommt ein sogenanntes Trainingsverfahren (Bishop [12], S. 2) zum Einsatz. Für dieses müssen die verwendeten Beispiele mit zusätzlichen Annotationen versehen sein. Hierzu gehört insbesondere die Klassenzuordnung, das sogenannte Label. Je nach Trainingsverfahren werden noch zusätzliche Informationen benötigt, wie bspw. eine Bounding-Box, die den Bereich von Interesse (engl. region of interest) oder die Koordinaten bestimmter Keypoints ausweist. Die im Training verwendeten Beispiele werden als Trainingsmenge bezeichnet. Diese muss Positiv- und Negativbeispiele für alle betrachteten Klassen beinhalten. Beim Trainingsverfahren wird über die Trainingsmenge iteriert. In jedem Schritt wird zunächst der Merkmalsdeskriptor extrahiert. Anschließend werden die Parameter des Klassifizierers entsprechend der Zielvorgabe des Beispiels angepasst. Das genaue Vorgehen hierbei hängt vom konkreten Klassifizierungsmodell und dem Trainingsverfahren ab.

Eine Schwierigkeit bei überwachten Lernalgorithmen besteht darin, dass die Regelmäßigkeiten der Trainingsdatenmenge erlernt werden müssen, und es dabei gleichzeitig gilt, diese für unbekannte Dateneingaben zu generalisieren. Ein Problem, das diesbezüglich auftreten kann, ist eine Überanpassung (engl. over-fitting) (Bishop [12], S. 6). Dies bedeutet, der Klassifizierer bildet die Trainingsdatenmenge zu genau ab und kann deshalb unbekannte Eingaben nicht zuordnen. Dieses Phänomen tritt im Allgemeinen bei einer geringen Diversität der Trainingsbeispiele bei einer gleichzeitig zu großen Anpassungsfähigkeit des Klassifizierermodells auf. Die Anpassungsfähigkeit (engl.: variance) eines Klassifizierers hängt vom Trainingsalgorithmus, dem Klassifizierungsmodell, sowie dessen Zahl an Konfigurationsparametern ab. Der umgekehrte Fall ist eine Unteranpassung. Diese tritt auf, wenn das Modell zu stark vereinfacht und hierdurch fehlerhafte Annahmen getroffen werden (engl.: bias). Das Problem, beide Fehlerquellen gleichzeitig zu minimieren, wird als Variance-Bias-Dilemma [27] bezeichnet. In gewissem Umfang lässt sich das Problem durch eine Verbesserung der Trainingsdaten oder durch eine Anpassung des Modells, insbesondere bezüglich der Anzahl an Konfigurationsparametern, adressieren.

Zu den bekanntesten Vertretern von Klassifizierungsmodellen, bei denen überwachte Trainingsverfahren eingesetzt werden, zählen die *Support Vector Machines* [28], *Künstliche Neuronale Netze* [29] oder auch die in Abschnitt 2.3.2.1 betrachteten *Random-Forest-Klassifizierer* [30]. Ein umfassender Überblick zu unterschiedlichen Modellen und Trainingsverfahren findet sich beispielsweise in [12].

Das unüberwachte Lernen unterscheidet sich vom überwachten Lernen dahingehend, dass die zu lernenden Klassen nicht vorgegeben werden. Entsprechend werden ungelabelte Beispieldaten verwendet. Ein entsprechendes Verfahren untersucht die Beispiele nach Gemeinsamkeiten und nimmt eine Klasseneinteilung selbstständig vor. Ein bekanntes Beispiele für ein unüberwachtes Lernverfahren ist die Clusteranalyse, die in Abschnitt 2.3.3 betrachtete wird.

2.3.2.1 *Random-Forest-Klassifizierer*

Das Random-Forest-Klassifizierermodell wurde im Jahr 2001 von Breiman [30] vorgeschlagen. Es basiert auf dem CART-Modell (Classification and Regression Trees) [31]. Bei diesem wächst ein binärer Entscheidungsbaum in der Trainingsphase angefangen beim Wurzelknoten nach und nach an. Die Trainingsbeispiele werden der Reihe nach an den Baum angelegt. Ein Knoten befindet sich dabei zunächst in der Evaluierungsphase. Hat er eine bestimmte Anzahl an Eingaben beobachtet, teilt er die bisherigen Beobachtungen anhand eines Schwellwerts und einer Merkmalsvektorkoordinate in zwei Teilmengen auf. Dabei wird die Koordinate gewählt, die die Varianzen der Teilmengen minimiert. Der Knoten erhält sodann zwei Nachfolgerknoten und befindet sich von da an in der Entscheidungsphase. Durch die Auswertung der gewählten Koordinate reicht er nun weitere Trainingsbeispiele an jeweils einen seiner beiden Nachfolgerknoten weiter. Nach Abschluss des Trainings erfolgt die Klassifizierung durch die Auswertung des Blattknotens, den die Beobachtung erreicht. Nachteil dieses Verfahrens ist, dass der Entscheidungsbaum dazu tendiert, asymmetrisch zu wachsen, und dass seine Struktur stark von der Reihenfolge der Trainingsbeispiele abhängt. Um dieses Problem zu vermeiden, werden beim Random-Forest-Modell, wie der Name suggeriert, mehrere Entscheidungsbäume eingesetzt, die jeweils auf einer kleineren, zufällig gewählten Teilmenge der Trainingsdaten trainieren. Dieses Vorgehen wird als *Bootstrap Aggregating* (kurz: Bagging) bezeichnet [32]. Ein weiterer Vorteil des Baggings ist, dass die Gefahr einer Überanpassung reduziert wird, da sich ein einzelner, tiefer Entscheidungsbaum stärker anpasst, als mehrere kleinere Bäume.

Ein zweites Prinzip, das beim Random-Forest-Modell angewendet wird, ist die *Random-Supspace-Methode* [33]. Bei dieser betrachten die einzelnen Entscheidungsbäume nicht den gesamten Merkmalsraum, sondern beschränken sich auf zufällig gewählte Koordinaten. Das Vorgehen hierbei ist wie folgt. Während der Evaluierungsphase beobachtet ein Knoten einen zufälligen Unterraum der ankommenden Merkmalsvektoren. Für jede Koordinate wird ein zufälliger Schwellwert innerhalb eines sinnvollen Wertebereichs gewählt. Nach einer bestimmten Anzahl von Beobachtungen wird nun diejenige Koordinate ausgewählt, mit der die Beobachtungen anhand des zufälligen Schwellwerts in zwei Teilmengen mit geringster Varianz unterteilt wird. Der Vorteil dieses Vorgehens ist,

dass durch die Beschränkung auf wenige, zufällige Koordinaten die Komplexität für die einzelnen Klassifizierer gesenkt wird. Trotz der Vereinfachung bei den einzelnen Bäumen werden sich Koordinaten, die starke Vorhersagen liefern, im Gesamtmodell durchsetzen, da diese von vielen Knoten unterschiedlicher Bäume ausgewählt werden und somit korrelieren. Die Komplexität wird also reduziert, ohne dabei die Gefahr einer Unteranpassung bei sehr hochdimensionalen Merkmalsdeskriptoren übermäßig zu erhöhen. Auf diese Weise kann mit hochdimensionalen Merkmalsvektoren effizient verfahren werden, wodurch auf eine vorgeschaltete Komplexitätsreduktion, bspw. eine PCA, verzichtet werden kann.

Bei der Klassifizierung ergeben sich die Wahrscheinlichkeiten für die Klassenzugehörigkeiten anhand der anteiligen Beobachtungen des Blattknotens, der bei der Auswertung eines Merkmalsvektors erreicht wird. Für das Gesamtergebnis werden die Ergebnisse aller Entscheidungsbäume gemittelt.

2.3.3 *Clusteranalyse und Vektorquantisierung*

Die *Clusteranalyse* ist eine Methode des unüberwachten Lernens zur statistischen Auswertung multivariater Datensätze. Das Ziel ist es, die Datenmenge in mehrere Untergruppen ähnlicher Elemente, sogenannte Cluster einzuteilen. Typischerweise wird eine Clusteranalyse eingesetzt, um Zusammenhänge in großen, unübersichtlichen Datenmengen zu identifizieren, die nicht offensichtlich bzw. nicht von vornherein bekannt sind. Für solche Untersuchungen hat sich der Begriff *Data-Mining* etabliert.

Eine ausführliche Einführung und Übersicht zu Clusteranalyse-Verfahren findet sich in [34]. Da das Finden eines globalen Optimums NP-schwierig ist, werden in der Praxis Näherungsverfahren eingesetzt. Zu den bekanntesten Ansätzen gehört das k-Means-Clustering und hierzu insbesondere der Llyod-Algorithmus [35]. Es handelt sich hierbei um ein iteratives Verfahren, welches eine festgelegte Anzahl von k Clustern ermittelt. Initial werden k Elemente zufällig aus dem Datensatz gewählt. Diese stellen die vorläufigen Zentren der Cluster dar, die als Zentroiden bezeichnet werden. Diese gilt es nun schrittweise zu optimieren. Dazu werden bei jeder Iteration zunächst alle Elemente ihrem am nächsten liegenden Zentroiden zugeordnet. Anschließend werden die Zentroiden so verschoben, dass sie jeweils im Zentrum ihrer Gruppe liegen. Durch Wiederholung dieses Vorgangs konvergiert der Algorithmus schließlich gegen ein lokales Optimum.

Im Rahmen der Mustererkennung kann eine Clusteranalyse (Bishop [12], S. 423 ff.) eingesetzt werden, um eine Stichprobe von Merkmalsdeskriptoren in Gruppen ähnlicher Merkmale aufzuteilen. Auf dieser Basis kann dann die Zuordnung weiterer Deskriptoren durchgeführt werden. Die Annahme, die dabei getroffen wird, ist, dass die Deskriptoren keiner kontinuierlichen Verteilung im Merkmalsraum folgen. Vielmehr wird davon ausgegangen, dass sich auf Grund

von Regelmäßigkeiten in der Datenmenge bestimmte Motive häufiger wiederholen und sich deshalb bei einer größeren Zahl beobachteter Merkmalsdeskriptoren in bestimmten Unterräumen lokale Häufungen bilden. Aufbauend auf dieser Annahme wird zunächst in einem Vorverarbeitungsschritt auf einer nicht zu geringen Menge stichprobenartig ermittelter Merkmalsdeskriptoren eine Clusteranalyse durchgeführt. Jeder identifizierte Cluster stellt dann eine lokale Merkmalshäufung dar, die durch einen Zentroiden repräsentiert wird. Die Zuordnung eines neuen Merkmalsdeskriptors erfolgt dann mittels der sogenannten Vektorquantisierung [36]. Die Zentroiden bilden für diese den Wertevorrat. Dieser wird mitunter auch als Codebuch oder Vokabular bezeichnet. Es wird nun der Zentroid mit dem geringsten Abstand zum Deskriptor ermittelt. Hierfür wird ein Verfahren zur Suche des nächsten Nachbarn (engl. Nearest-Neighbour-Search; Bishop [12], S. 124 ff.) benötigt. Eine lineare Suche würde eine Laufzeit von $O(Kd)$ beanspruchen, mit K als der Anzahl der Zentroiden und d der Dimension des Merkmalsraums. Da dies für praktischen Anwendungen zu langsam ist, werden effizientere Methoden bspw. auf Basis einer Raumaufteilung (engl.: space partitioning) eingesetzt. Eine Einführung zu diesem Thema findet sich hier [37]. Eine zur Raumaufteilung häufig verwendete Datenstruktur ist der *k-d-Baum* (Grauman und Leibe [2], S. 32). Es handelt sich um einen binären Suchbaum, bei dem jeder Knoten ein Element des Vokabulars repräsentiert. Auf Höhe des Elements wird eine Hyperebene aufgespannt, die den k-dimensionalen Suchraum senkrecht zu einer der Dimensionen in zwei Halbräume teilt. Aufgabe eines Konstruktionsalgorithmus ist es, einen möglichst ausgewogenen Suchbaum aufzuspannen. Die Suche des nächsten Nachbarn kann dann mittels einer Tiefensuche durchgeführt werden.

Für die Merkmalsbetrachtung stellt die Vektorquantisierung eine erhebliche Reduktion der Komplexität dar. Sie ermöglicht damit eine effizientere Verarbeitung hochdimensionaler Merkmalsdeskriptoren. Da die Vektorquantisierung unüberwacht und ohne den Einsatz von Modellwissen durchgeführt werden kann, stellt sie ein wichtiges Hilfsmittel für den Entwurf generischer Mustererkennungsverfahren dar.

3

OBJEKTERKENNUNG

Dieses Kapitel behandelt das Thema der Objekterkennung. In Abschnitt 3.1 wird zunächst auf die Definition der Problemstellung und der damit verbundenen Begrifflichkeiten eingegangen. Abschnitt 3.2 geht auf die Unterschiede zwischen Top-Down- und Bottom-Up-Ansätzen ein. In Abschnitt 3.3 werden verschiedene Strategien zu Objekterkennung erläutert. Darauf aufbauend wird in Abschnitt 3.4 der Bag-of-Features-Ansatz näher betrachtet, der für die in Kapitel 6 beschriebene Konzeption und Umsetzung des praktischen Teils dieser Arbeit eine wichtige Rolle spielt. In Abschnitt 3.5 werden Strategien zur effizienten Objektlokalisierung betrachtet. Die Betrachtung dient dabei der Motivation des aufmerksamkeitsbasierten Ansatzes.

3.1 PROBLEMSTELLUNG UND BEGRIFFE

Bei der Objekterkennung (Grauman und Leibe [2], S. 1 ff.) geht es darum, Objekte in einem Bild oder einer Videosequenz zu identifizieren. In der Regel werden hierbei natürliche Bilder betrachtet. Der Begriff *natürlich* besagt, dass ein Bild in einem für den Menschen typischen Umfeld mit einem im optischen Spektrum arbeitenden Kamerasensor aufgenommen wird. Dies ist als Abgrenzung zu speziellen Aufnahmetechniken (Gonzalez und Woods [11], S. 7 ff.) wie Wärmebildern und ähnlichem zu verstehen, die hier nicht betrachtet werden. Unter einem Objekt ist im engeren Sinne zunächst ein Gegenstand zu verstehen. Für eine enge Auslegung des Begriffes besteht aber keine Notwendigkeit. Im weitesten Sinne kann er sich auf alles beziehen, was eine in sich geschlossene visuelle Darstellung aufweist. Unter einer Objektklasse oder -kategorie ist eine Gruppe von Objekten mit ähnlichen visuellen Eigenschaften zu verstehen. Objekte, die einer Kategorie angehören, werden im Folgenden als *Instanz* dieser Kategorie bezeichnet.

Bei der Formulierung von Objekterkennungsproblemen werden in der Literatur Begriffe wie Erkennung, Detektion, Lokalisierung etc. nicht unbedingt einheitlich verwendet. In dieser Arbeit wird den Definitionen nach Perona [38] gefolgt. Hiernach wird zunächst zwischen einem komplexen Bild und einer Bildkachel (engl. patch) unterschieden. Eine Kachel kann entweder ein Bild für sich oder den Ausschnitt eines größeren Bildes darstellen. Bei dem Problem der *Verifikation* ist zu entscheiden, ob eine Bildkachel eine Instanz einer bestimmten Objektkategorie zeigt oder nicht. Als Nebenbedingung können hierbei Anforderungen hinsichtlich der Genauigkeit gestellt werden. Es kann gefordert werden, dass ein

positiver Befund nur dann zu stellen ist, wenn die Kachel das Objekt möglichst genau einrahmt, ohne zu viel Hintergrund (engl.: clutter) zu zeigen, oder zu viel von dem Objekt abzuschneiden (siehe hierzu Abschnitt 7.2). Das Problem der *Klassifizierung* ist eine Erweiterung der *Verifikation*, bei der mehrere Objektklassen zur Auswahl stehen. Beim *Präsenzproblem* gilt es zu entscheiden, ob ein komplexes Bild mindestens eine Instanz einer bestimmten Objektkategorie zeigt oder nicht. Die Größe des Objekts und seine Lage im Bild muss dabei nicht ermittelt werden. Hierin unterscheidet es sich vom Problem der *Lokalisierung*, bei dem die Position und die Größe für alle Objektinstanzen ermittelt werden soll. Dies geschieht in der Regel durch die Angabe von Bounding-Boxen. Werden mehrere Objektkategorien betrachtet, müssen zusätzlich auch noch die Label der Instanzen angegeben werden. In dieser Arbeit wird das Problem der Lokalisierung betrachtet. Wird im weiteren Verlauf dieser Arbeit also allgemein von Objektdetektion oder -erkennung gesprochen, geschieht dies im Hinblick auf diese Problemstellung. Die Problemstellungen der Verifikation und Klassifizierung sind dabei insofern von Bedeutung, als dass sie als Teilprobleme des Lokalisierungsproblems aufgefasst werden können, die nur einen von vielen möglichen Bildausschnitten betrachten. Dem Präsenzproblem kommt hier nur insoweit Bedeutung zu, als dass es als eine Vereinfachung des Lokalisierungsproblems aufgefasst werden kann. Der Vollständigkeit halber sei noch die Problemstellung des *Verstehens* erwähnt, welche in dieser Arbeit jedoch nicht betrachtet wird. Hierbei handelt es sich um eine Erweiterung des Lokalisierungsproblems, bei der zusätzlich die Interaktion und die Beziehung der erkannten Objekte zueinander beschrieben werden soll.

3.2 TOP-DOWN VERSUS BOTTOM-UP

Bei der Objekterkennung kann eine generelle Unterscheidung zwischen sogenannten *Top-Down-* und *Bottom-Up*-Verfahren getroffen werden. Der Unterschied besteht darin, dass beim Top-Down-Ansatz gezielt Modellwissen bezüglich bestimmter Objektkategorien eingesetzt wird, während beim Bottom-Up-Ansatz Eigenschaften betrachtet werden, die sich ganz allgemein auf Objekte anwenden lassen und somit kein spezifisches Vorwissen erfordern.

Bei allen Problemstellungen, die im letzten Abschnitt beschrieben wurden, werden Objektkategorien betrachtet. Ansätze zur Lösung solcher Problemstellungen nutzen in der Regel Modellwissen über die betrachteten Objektkategorien und sind somit den Top-Down-Ansätzen zuzurechnen. Das typische Vorgehen beim Top-Down-Ansatz ist es, das benötigte Modellwissen anhand von Trainingsbeispielen zu erlernen, welches dann bei der Objekterkennung eingesetzt werden kann (Grauman und Leibe [2], S. 3). Werden die gelernten Informationen in unbekannten Dateneingaben in ähnlicher Form wiedererkannt, kann auf ein entsprechendes Objekt geschlossen werden. Für dieses Vorgehen werden Metho-

den zur Merkmalsextraktion, zum maschinellen Lernen und zur Klassifizierung eingesetzt, wie sie in Kapitel 2 beschrieben wurden. Ein wichtiger Aspekt bei Top-Down-Verfahren ist es, in welcher Form die Trainingsbeispiele vorliegen müssen. Im einfachsten Fall besteht ein Trainingsbeispiel aus einem Bild mit einem Klassenlabel. Dies ist bspw. bei dem in Abschnitt 3.4 betrachteten Bag-Of-Features-Modell der Fall. Bei komplexeren Modellen können auch weitere Informationen herangezogen werden, wie bspw. die Lage bestimmter Objektteile. Zwar versprechen komplexere Modelle bessere Detektionsraten, jedoch erfordert das Erstellen von Trainingsbeispielen im Allgemeinen manuelle Arbeit und ist recht zeitaufwändig. Insbesondere wenn eine größere Anzahl unterschiedlicher Objektkategorien betrachtet werden soll, haben einfachere Modelle hier einen entscheidenden Vorteil.

Der Bottom-Up-Ansatz (bspw. [39]) nähert sich dem Problem der Objekterkennung von einer anderen Seite. Es wird kein spezifisches Modellwissen eingesetzt, sondern es werden Eigenschaften betrachtet, die unabhängig von der Zugehörigkeit zu einer bestimmen Kategorie für Objekte im Allgemeinen gelten. Beispielsweise kann die Annahme getroffen werden, dass Objekte eine geschlossene Kontur aufweisen, die sie von ihrer Umgebung abgrenzen. Entsprechend gilt es bei Bottum-Up-Ansätzen, Objekte in einem Bild zu detektieren, ohne dieses dabei zu kategorisieren. Der Vorteil von Bottom-Up-Verfahren ist die generische Einsetzbarkeit, die durch den Verzicht auf spezifisches Modellwissen gegeben ist.

Wie bereits erwähnt wurde, zielen konkrete Problemstellungen häufig auf die Kategorisierung von Objekten ab, was durch Bottom-Up-Verfahren nicht geleistet wird. Zur Lokalisierung von Objekten ist es jedoch auch möglich, Bottom-Up- und Top-Down-Methoden zu kombinieren. Die Idee dabei ist, dass ein Bottom-Up-Detektor zunächst die Lage der Objekte im Bild ermittelt. Diese werden anschließend von einem Klassifizierer einer Objektkategorie zugeordnet. Dieses Vorgehen entspricht genau dem in Kapitel 1.2 skizzierten Vorgehen, wobei dort die Aufgabe der Bottom-Up-Detektion von einem Salienzdetektor übernommen wird.

3.3 STRATEGIEN ZUR OBJEKTKLASSIFIZIERUNG

Dieser Abschnitt erläutert einige wichtige Strategien zur Objektklassifizierung, die für die thematische Einordnung der in dieser Arbeit präsentierten Methoden relevant sind. Für eine umfassende Übersicht zu verschiedenen Objekterkennungsverfahren und die Entwicklung der letzten fünfzig Jahre hierzu sei an dieser Stelle auf Andreopoulos und Tsotsos [40] verwiesen.

Die Klassifizierung eines Objekts erfordert die Betrachtung seiner verschiedenen Merkmale sowie deren örtlichen Beziehungen zueinander. Wie Abbildung 3.3.1 zeigt, können hierbei verschiedene Strategien verfolgt werden. Ansatz (a), der als holistischer Ansatz (Grauman und Leibe [2], S. 8) bezeichnet wird, stellt

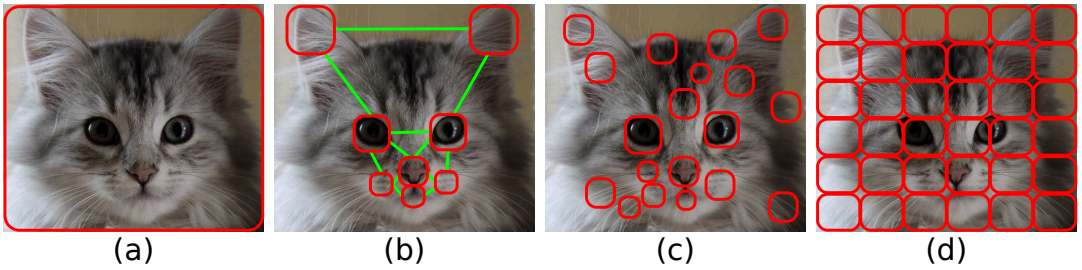

Abbildung 3.3.1: Strategien zur Merkmalsextraktion. (a) Holistisch. (b) Teilebasiert. (c) Keypoint-Detektion. (d) Dichtes Raster.

den einfachsten Fall dar. Hierbei wird nach einer starren Vorschrift lediglich ein einziger Merkmalsvektor extrahiert, der dann den Inhalt der gesamten Kachel repräsentiert. Bei dem teilebasierten Ansatz (b) (Grauman und Leibe [2], S. 69 ff.) werden Merkmalsvektoren für wichtige Punkte bzw. Teile des Objekts gesucht und extrahiert. Jeder Merkmalsvektor beschreibt dann einen kleineren Ausschnitt der gesamten Kachel. Die Objektteile werden im Vorfeld durch den Einsatz von Modellwissen festgelegt. Die Schwierigkeit besteht darin, dass die Objektteile sich von Bild zu Bild an abweichenden Positionen befinden können. Zur Laufzeit müssen sie deshalb durch ein geeignetes Verfahren lokalisiert werden bzw. ihre Abwesenheit muss festgestellt werden können. Als typisches Beispiel für Ansatz (b) sei das *Bunch-Graph-Matching* [41] genannt, bei dem das beschriebene Vorgehen mit Hilfe einer geometrischen Graphenstruktur realisiert wird, was in der Abbildung durch die zusätzlichen grünen Verbindungslinien dargestellt wird. Während der Graph zur Bestimmung der ungefähren Lage der lokalen Merkmale dient, erfolgt die exakte Lokalisierung durch der Auswertung des Phasenspektrums. Der Vorteil im Vergleich zur Variante (a) besteht darin, dass gezielt die für eine Objektkategorie relevanten Merkmale betrachtet werden. Der Nachteil hingegen ist, dass der Einsatz von Modellwissen eine generische Einsetzbarkeit faktisch ausschließt. Es kann lediglich eine Objektkategorie betrachtet werden. Hinzu kommt, dass für die Umsetzung eine Menge von Beispielbildern erforderlich ist, bei denen die Koordinaten der Teile bekannt sein müssen. Diese müssen in der Regel manuell oder zumindest halbautomatisch unter Zuhilfenahme eines lernenden Assistenzsystems annotiert werden, was einen entsprechenden Arbeitsaufwand bedeutet. Bei den Ansätzen (c) und (d) (Grauman und Leibe [2], S. 66) wird hingegen kein Modellwissen eingesetzt, was eine generische Anwendung erlaubt. Bei Ansatz (c) werden Keypoints mittels eines so genannten Keypoint-Detektors automatisch lokalisiert. Ein Beispiel hierzu wurde bereits in Abschnitt 2.1.2 mit dem SIFT-Verfahren gegeben. Der Keypoint-Detektor lokalisiert Punkte, die sich aufgrund besonderer Merkmale bei verschiedenen Instanzen einer Objektkategorie mit einer gewissen Wahrscheinlichkeit wiederfinden lassen (Grauman und Leibe [2], S. 13 ff.). Hierbei können auch die Skalierung und Orientierung eines Punktes ermittelt werden, was wie-

derum beim Merkmalsabgleich berücksichtigt werden kann, um die Robustheit gegenüber diesen Faktoren zu erhöhen. Bei Ansatz (d) werden die Merkmale auf einem dichten Raster extrahiert. Für diesen Ansatz lassen sich ebenfalls die SIFT-Deskriptoren verwenden, die in Abschnitt 2.3.1.3 betrachtet wurden. Diese werden dann bei fester Skalierung und Rotation auf einem entsprechenden Raster extrahiert. Diese Vorgehensweise wird als *DSIFT* (dense SIFT) [42] bezeichnet.

Bei allen hier betrachteten Ansätzen muss sich der Merkmalsextraktion ein Verfahren anschließen, welches die Merkmale hinsichtlich einer Klassenzuordnung auswertet. Hierzu wurde bereits das Thema der Klassifizierung in Abschnitt 2.3.2 betrachtet. Bei Ansatz (a) kann der holistische Merkmalsvektor direkt von einem Klassifizierer ausgewertet werden. Bei Ansatz (b) kann für jeden der lokalen Merkmalsvektoren ein eigener Klassifizierer verwendet werden, der dann als lokaler Experte bezeichnet wird. Das Gesamtergebnis kann dann beispielsweise durch einen Mehrheitsentscheid bestimmt werden. Bei den Ansätzen (c) und (d) kommen Votierungssysteme (Grauman und Leibe [2], S. 83 ff.) in Frage, bei denen nach einer signifikanten Anzahl an Merkmalsdeskriptoren gesucht wird, die für ein bestimmtes Ergebnis sprechen (bspw. [43]). Eine andere Vorgehensweise ist es, alle extrahierten Deskriptoren nach einer bestimmten Vorschrift zu einem einzigen Deskriptor zusammenzufassen. Man spricht dann in diesem Zusammenhang von *lokalen* und *globalen* Deskriptoren (Grauman und Leibe [2], S. 7 ff.). Ein solcher globaler Deskriptor kann dann wiederum durch einen Klassifizierer ausgewertet werden. Dieser Ansatz wird beim Bag-Of-Features-Modell verfolgt, das im nächsten Abschnitt betrachtet wird.

3.4 DAS BAG-OF-FEATURES-MODELL

Beim *Bag-Of-Features-Modell* (BoF) [44] handelt es sich um einen leitsungsfähigen Ansatz zur Klassifizierung visueller Daten. Er basiert auf der Bestimmung der Häufigkeitsverteilung visueller Merkmale. Zur Ermittlung der Häufigkeitsverteilung werden zunächst über die Bildfläche verteilt Merkmalsvektoren extrahiert. Diese werden anschließend mittels Vektorquantisierung quantisiert (siehe Abschnitt 2.3.3). Anschließend wird über die quantisierten Merkmale ein Histogramm erstellt (siehe Abschnitt 2.3.1.1). Jeder Codebucheintrag entspricht dabei einem Histogrammkanal. Dieses Prinzip hat seine Ursprünge zum einen im *Bag-Of-Words-Modell* [45], bei dem auf die gleiche Weise die Häufigkeitsverteilung von Wörtern einer natürlichen Sprache zur Klassifizierung von Dokumenten eingesetzt wird, und zum anderen in der Texturerkennung, bei der Codebücher bestehend aus Texturelementen zur Klassifizierung von Texturmustern eingesetzt werden [46]. Im Folgenden wird in Abschnitt 3.4.1 der Aufbau eines Trainings- und Klassifizierungsverfahrens auf der Basis des Bag-Of-Features-Modells betrachtet. In Abschnitt 3.4.2 werden Erweiterungen zu diesem Grundprinzip betrachtet.

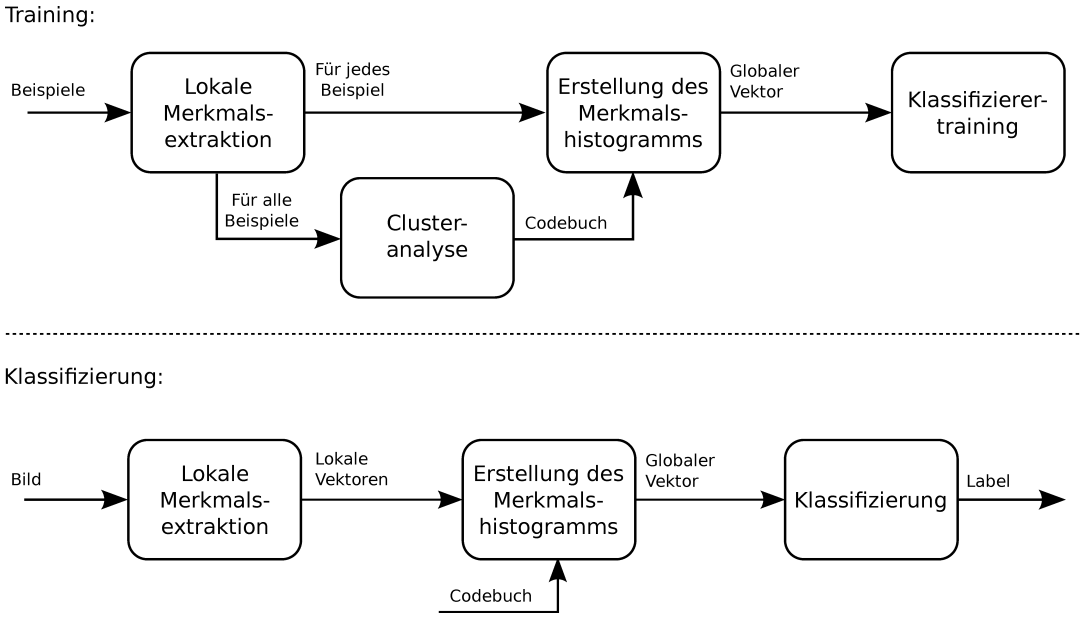

Abbildung 3.4.1: Der Bag-of-Features-Ansatz.

3.4.1 *Training und Klassifizierung*

Abbildung 3.4.1 zeigt das Vorgehen für ein Trainings- und Klassifizierungsverfahren, dass auf dem BoF-Prinzip aufbaut (vgl. Grauman und Leibe [2], S. 109 f.). Für das Trainingsverfahren wird eine Menge von positiven und negativen Bildbeispielen für alle betrachteten Objektklassen benötigt. Diese müssen mit einem entsprechenden Klassenlabel versehen sein. Weitere Annotationen werden nicht benötigt. Im ersten Schritt erfolgt die Merkmalsextraktion. Das BoF-Prinzip ist an kein bestimmtes Extraktionsverfahren gebunden. Bei natürlichen Bildern stellen zum Beispiel die in Abschnitt 2.3.1.3 beschriebenen SIFT-Merkmale eine gute Wahl dar. Diese können gleichmäßig und dicht verteilt über die Bildfläche (DSIFT) oder auf Basis eines Keypoint-Detektors extrahiert werden (vgl. Abschnitt 3.3). In [47] wurden hierzu verschiedene Sampling-Strategien verglichen. Die Untersuchung zeigte, dass sich die Ergebnisse mit der Anzahl der Merkmalsvektoren verbessern, eine dichte Merkmalsextraktion folglich zu besseren Resultaten führt, als eine auf Keypoints basierende Strategie.

Beim Trainingsverfahren wird die Merkmalsextraktion zunächst auf der gesamten Trainingsmenge durchgeführt, um über einen ausreichend großen Merkmalspool als Basis für das Vektorquantisierungsverfahren zu verfügen. Für dieses kann eine Clusteranalyse eingesetzt werden (siehe Abschnitt 2.3.3). Wird das k-Means-Verfahren eingesetzt, so bestimmt k die Anzahl der Codebucheinträge und somit die Größe des globalen Merkmalsvektors. Um diesen zu erstellen, werden zunächst die lokalen Merkmale extrahiert, anschließend quantisiert und in ein k-dimensionales Histogramm einsortiert. Die globalen Vektoren, die aus

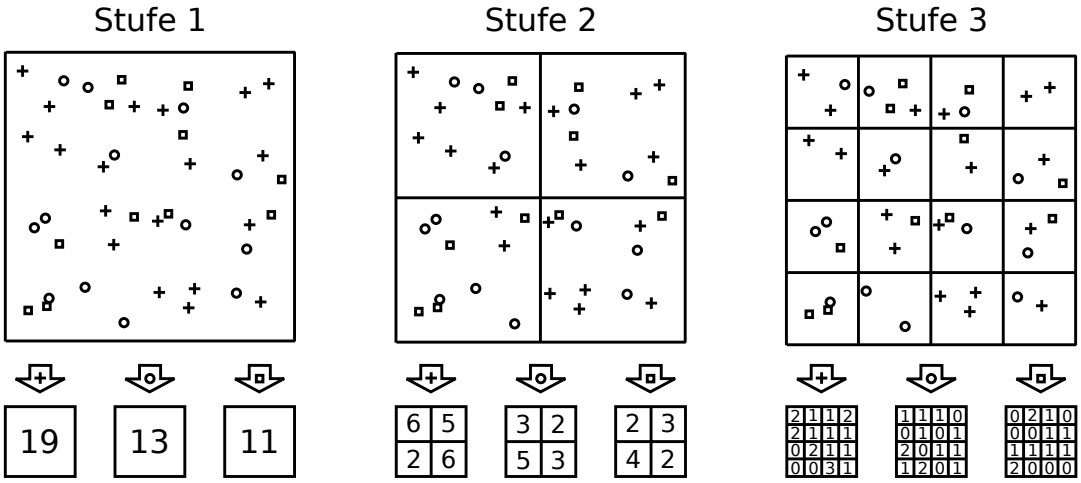

Abbildung 3.4.2: Beispiel für eine dreistufige Pyramide und einem Codebuch mit drei Einträgen nach Grauman und Darrell [48].

den einzelnen Bildbeispielen extrahiert werden, stellen die Trainingsmenge für das Training des Klassifizierers dar, wie es in Abschnitt 2.3.2 beschrieben wurde. Beim Klassifizierungsverfahren laufen die Merkmalsextraktion und die Erstellung des globalen Merkmalsvektors auf die gleiche Weise ab. Die Klassifizierung des globalen Merkmalsvektors erfolgt dann im Anschluss.

3.4.2 *Erweiterungen*

Wie bereits in Abschnitt 2.3.1.1 erläutert wurde, ist ein Nachteil der Histogrammdarstellung der Verlust der örtlichen Informationen. Dies gilt entsprechend auch für den Bag-Of-Features-Ansatz. Die reine Häufigkeitsverteilung der lokalen Merkmale sagt nichts über ihre Lage zueinander aus und kann somit die globale Struktur des Objekts nicht beschreiben. Lazebnik et al. [49] schlagen eine Erweiterung vor, bei der die Anordnung der lokalen Merkmale berücksichtigt wird. Hierzu wird die eindimensionale Histogrammdarstellung durch die Pyramidendarstellung nach Grauman und Darrell [48] ersetzt, die auf den Ortsbereich des Bildes angewendet wird. Abbildung 3.5.1 zeigt das Vorgehen anhand eines Beispiels für ein Codebuch mit drei Einträgen. Das Bild wird stufenweise in ein immer feiner werdendes Raster aufgeteilt. Für jede Zelle wird ein eigenes Histogramm erstellt. Der globale Merkmalsvektor kann schließlich durch einfaches Konkatenieren der Histogramme gebildet werden.

Eine andere Herangehensweise findet sich bei Grzeszick et al. [50]. Auch hier wird der Ortsbereich des Bildes in Zellen unterteilt. Die Zugehörigkeit zu einer Zelle wird aber im lokalen Merkmalsvektor codiert, sodass ein Teil des Vektors die Erscheinung und ein anderer Teil die Lage beschreibt. Hierdurch

wird erreicht, dass sich bei der Clusteranalyse für die einzelnen Bereiche Cluster für Merkmalsvektoren mit ähnlicher Erscheinung bilden.

3.5 STRATEGIEN ZUR EFFIZIENTEN OBJEKTLOKALISIERUNG

In den vorangegangenen Abschnitten wurden Strategien und Verfahren zur Objektklassifizierung erläutert, die entsprechend den Problemdefinitionen aus Abschnitt 3.1 auf das Verifizierungs- und Klassifizierungsproblem angewendet werden können. Bei der Problemstellung der Lokalisierung kommt erschwerend hinzu, dass die Größe und Position der Objekte bestimmt werden muss. Klassische Detektionsmethoden verwenden hierfür in der Regel eine einfaches Suchmuster. Ein Beispiel hierfür ist der Einsatz eines *gleitenden Fensters* (Grauman und Leibe [2], S. 78 f.), wie es in Abbildung 3.5.1 dargestellt ist. Dieses durchläuft auf einem festgelegten Raster alle möglichen Bildpositionen. An jeder Position wird dann das Verfahren zu Objektklassifizierung durchgeführt, um festzustellen, ob sich ein gesuchtes Objekt innerhalb des Fensters befindet oder nicht. Um Objekte unterschiedlicher Größe lokalisieren zu können, wird der Vorgang auf verschiedenen Skalierungsstufen wiederholt. Dieses vergleichsweise einfache Vorgehen hat zwei wesentliche Nachteile. Die hohe Zahl der Positionen, die es zu überprüfen gilt, macht das Vorgehen sehr zeitaufwändig. Des Weiteren besteht an jeder Position die Gefahr einer Fehldetektion, was insbesondere bei der Betrachtung vieler Objektkategorien zu einer hohen Fehlerrate führen kann. Im Folgenden werden deshalb einige Strategien zur Verbesserung des Vorgehens betrachtet.

Eine Möglichkeit zur Reduktion von Fehldetektionen besteht in einer Verbesserung der Klassifzierung durch die Auswertung von Umgebungsinformationen. Allein durch die Betrachtung des Umfeldes lassen sich Aussagen über das mögliche Vorhandensein eines Objektes treffen, ohne dabei die Merkmale des Objekts selbst betrachten zu müssen. Ein Beispiel hierzu auf Basis einer Szenenerkennung findet sich bei Torralba et al. [51].

Eine naheliegende Strategie zur reinen Beschleunigung der Lokalisierung besteht in einer umfassenden Parallelisierung der Berechnungen durch den Einsatz geeigneter Hardware. Ein erhöhter Hardwareaufwand bedeutet jedoch auch höhere Herstellungskosten, mehr Platzbedarf und ein erhöhter Energieverbrauch zur Laufzeit, was insbesondere bei der Entwicklung von Systemen mit begrenzten Energiereserven oder begrenzten Möglichkeiten zur Wärmeabfuhr problematisch ist.

Eine mögliche Strategie zur Beschleunigung der Lokalisierung besteht in der Reduktion von Redundanzen, die durch eine Überlappung des Suchfensters an benachbarten Positionen entstehen. Ja nach Klassifizierungsverfahren kann dies ausgenutzt werden, um bereits durchgeführte Berechnungen wiederzuverwenden bzw. Berechnungen gleichzeitig für mehrere Positionen durchzuführen. Im

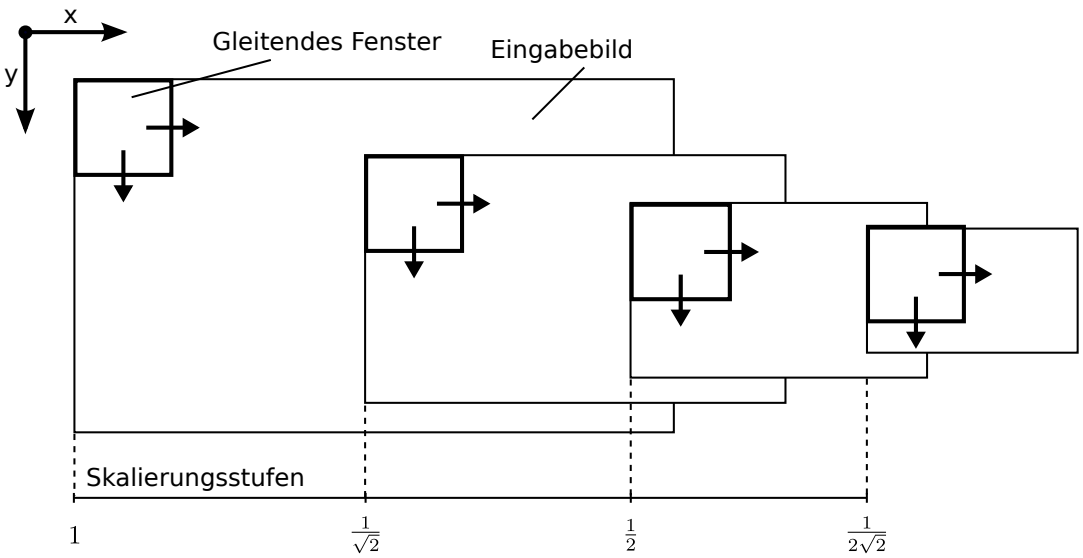

Abbildung 3.5.1: Gleitendes Fenster für unterschiedliche Objektgrößen.

Allgemeinen ist dies kein triviales Unterfangen und erfordert gewisse Vorkehrungen bei der Modellierung des Klassifizierers. Ein Beispiel hierzu stellen die Faltungsnetzwerke nach Osadchy et al. [52] dar. Es handelt sich dabei um ein künstliches neuronales Netz, dessen Neuronen wie Faltungsmasken modelliert sind. Werden diese Masken direkt auf das gesamte Eingabebild angewendet, kann durch Berechnungen im Fourier-Raum erhebliche Rechenzeit eingespart werden [53].

Eine weitere Strategie ist es, die Robustheit des Klassifizierungsverfahrens gegenüber affinen Transformationen zu erhöhen. Hierdurch wird es möglich, ein gröberes Suchrasters zu verwenden. Wie in Abschnitt 2.3.1 beschrieben wurde, kann das Thema der Robustheit bei der Merkmalsbetrachtung adressiert werden. Eine gängige Vorgehensweise ist die Vernachlässigung der örtlichen Informationen durch die Bildung von Histogrammen. Eine andere Möglichkeit ist die explizite Lokalisierung bestimmter örtlicher Merkmale (vgl. Abschnitt 3.3). Eine weitere Möglichkeit besteht darin, die Variationen innerhalb des Trainingsdatensatz zu erhöhen, wie in Abbildung 3.5.2 veranschaulicht wird. Der Klassifizierer lernt dann im Trainingsverfahren die unterschiedlichen Variationen zu erkennen. Der Nachteil daran ist, dass eine Erhöhung der Variation einen komplexeren Klassifizierer erfordert, um die Gefahr einer Unteranpassung zu vermeiden. In jedem Fall gilt, das die Robustheit gegenüber affinen Transformationen nur innerhalb bestimmter Grenzen möglich ist. Des Weiteren wird durch die Verwendung eines groben Suchrasters die Präzision der Lokalisierung reduziert.

Eine weitere Optimierungsstrategie ist die Verwendung eine Kaskade schwacher Klassifizierer, wie sie von Viola und Jones [54] vorgeschlagen wird. Das Prinzip ist in Abbildung 3.5.3 dargestellt. Es wird die Tatsache ausgenutzt, dass bei der Suche nach Objekten das Nichtvorliegen eines Objekts das wesentlich häufigere

Abbildung 3.5.2: Beispiel zur Erhöhung der Variationen im Trainingsdatensatz. Vier unterschiedliche affine Transformationen des gleichen Trainingsbildes.

Ereignis ist. Jeder der schwachen Klassifizierer kann mit einer hohen Bestimmtheit das Nichtvorliegen eines Objektes erkennen und die Eingabe entsprechend verwerfen. Das Vorliegen eines Objektes gilt hingegen erst als zuverlässig erkannt, wenn die gesamte Kaskade mit positiven Befunden durchlaufen wurde. In der Praxis werden viele der negativen Eingaben bereits in einer frühen Stufe verworfen, was sich günstig auf die Laufzeit auswirkt. Ein schwacher Klassifizierer kann aus dieser Betrachtung heraus als Vorfilter betrachtet werden. Der Nachteil dieses Ansatzes ist, dass das Trainingsverfahren recht zeitaufwändig ist. Des Weiteren lässt sich das Verfahren nicht ohne Weiteres auf mehrere Objektklassen erweitern. Ein ähnlicher Ansatz besteht darin, die Suche mittels gleitendem Fenster zunächst mit einem Klassifizierer geringer Komplexität durchzuführen. Anschließend werden die besten Treffer mit einem komplexeren Klassifizierer untersucht. Auch dieser Ansatz macht sich zu Nutze, dass die Anzahl der Objekte im Verhältnis zu den untersuchten Teilbildern eher gering ist. Ein Beispiel hierfür findet sich bei Vedaldi et al. [55]. Dort wird ein dreistufiger Prozess mit einem linearen, einem quasi-linearen und einem nicht-linearen Klassifizierer vorgeschlagen.

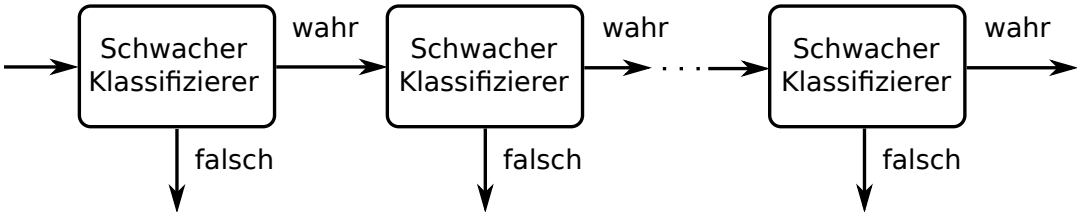

Abbildung 3.5.3: Kaskade schwacher Klassifizierer nach Viola und Jones [54].

Eine weitere Möglichkeit ist der Einsatz von Bottom-Up-Bildsegmentierung. Ein entsprechendes Beispiel hierzu findet sich bei van de Sande et al. [56]. Das Bild kann im ersten Schritt in Segmente aufgeteilt werden, die dann im zweiten Schritt klassifiziert werden. Auf diese Weise kann der Suchraum stark eingeschränkt werden. Der Nachteil daran ist, dass nur solche Objekte detektiert werden können, die zuvor auch hinreichend genau segementiert werden konnten.

Es handelt sich hierbei im Wesentlichen um eine Variante der in Abschnitt 3.2 erörterten Kombination einer Bottom-Up- und einer Top-Down-Methode. Der aufmerksamkeitsbasierte Ansatz, der in dieser Arbeit betrachtet wird, fällt ebenfalls in diese Kategorie. Der Salienzdetektor schränkt den Suchraum auf vergleichsweise wenige Kandidaten ein, die dann mittels eines Klassifizierers genauer untersucht werden. Dieser Ansatz basiert auf die in Kapitel 1.2 hervorgehobene These, dass interessante Objekte salient sind. Der Vorteil dieses Ansatzes besteht darin, dass ein Salienzdetektor generisch in Bezug auf die Objektkategorie eingesetzt werden kann. Im nun folgenden Kapitel werden die Grundlagen hierzu betrachtet.

4

VISUELLE AUFMERKSAMKEIT

Das visuelle System des Menschen ist überaus leistungsfähig (Jenkin und Harris [1], S. 3 f.). Mit mehr als 132 Millionen Photorezeptoren der Retina und etwa einer Millionen Nervenfasern des Sehnervs verarbeitet es fortlaufend eine sehr große Informationsmenge. Erstaunlich daran ist, dass der Mensch diese Informationsmenge effektive verarbeiten und dabei komplexe Aufgaben wie beispielsweise das Erkennen von Objekten und Personen problemlos erledigen kann. Wie jedoch ist der Mensch in der Lage, eine solche Informationsfülle scheinbar mühelos zu bewältigen? Es lässt sich vermuten, dass das Gehirn alle visuellen Informationen parallel verarbeitet. Diese Vermutung drängt sich aufgrund der neuronalen Netzwerkstruktur des Gehirns auf, die grundsätzlich eine parallele Informationsverarbeitung erlaubt. Sie ist jedoch falsch. Tatsächlich arbeitet das visuelle System des Menschen wesentlich ressourcenschonender. Eine wichtige Rolle zum Verständnis hierzu spielt das Prinzip der visuellen Aufmerksamkeit (Jenkin und Harris [1], S. 1 ff.). Das visuelle System hat die Eigenschaft, auf bestimmte Sehinhalte, von denen eine gewisse Stimulation ausgeht, mit erhöhter Aufmerksamkeit zu reagieren. Andere Sehinhalte treten hingegen in den Hintergrund bzw. werden ausgeblendet. Diese Art der Informationsfilterung geschieht fortwährend, schnell, unbewusst und mühelos (Jenkin und Harris [1], S. 93 ff.). Auf diese Weise wird der Fokus bereits auf potentiell wichtige bzw. interessante Sehinhalte gelegt, noch bevor sich deren Semantik dem Betrachter erschlossen hat. Erst im Anschluss an diese *präattentiven* Prozesse analysieren komplexere, *attentive* Prozesse, welche Informationen von den jeweiligen Reizen ausgehen. Es handelt sich bei dem Prinzip der Aufmerksamkeit um eine grundlegende Eigenschaft der menschlichen Wahrnehmung, die eine effiziente Informationsauswertung überhaupt erst möglich macht. Es ist das Bestreben aktueller Forschung, diese präattentiven Prozesse besser zu verstehen. Die Erkenntnisse, die aus dieser Forschung gewonnen wurden und noch werden, sind für technische Anwendungen interessant, um diese effizienter zu gestalten. Anwendungsmöglichkeiten können sich auf vielen Gebieten ergeben, bei denen visuelle Daten ausgewertet werden. Insbesondere für die visuelle Mustererkennung ist das Prinzip der Aufmerksamkeit interessant.

Im folgenden Abschnitt 4.1 werden Theorien zur visuellen Aufmerksamkeit des Menschen betrachtet. Abschnitt 4.2 geht auf die Möglichkeit ein, das Prinzip der visuellen Aufmerksamkeit auf mathematische Modelle zu übertragen. Eines der bekanntesten und einflussreichsten Modell dieser Art ist das von Itti et al.

[57], welches in Abschnitt 4.3 vorgestellt wird. Weitere Modelle aus der Literatur werden anschließend in Kapitel 5 betrachtet.

4.1 VISUELLE AUFMERKSAMKEIT BEIM MENSCHEN

Das visuelle System des Menschen dient aufgrund seiner Leistungsfähigkeit als Vorbild und Referenz für Forschungen auf dem Gebiet Computer-Vision. Auch wenn seine Funktionsweise bisher noch nicht vollständig verstanden ist, so sind dennoch viele Teilaspekte untersucht, aus denen sich Theorien und praktische Methoden ableiten lassen. Dies gilt auch für den Bereich der visuellen Aufmerksamkeit. Wichtige Erkenntnisse zur menschlichen visuellen Aufmerksamkeit, die bei der Ableitung eines mathematischen Aufmerksamkeitsmodells hilfreich sind, sollen im Folgenden erläutert werden.

Zunächst stellt sich die Frage nach einer grundlegenden Definition des Begriffs der Aufmerksamkeit. Ein bekannter Versuch hierzu findet sich bei dem als Vater der Psychologie bekannten William James [58]:

> „Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others [...]“

> „Jeder weiß, was Aufmerksamkeit ist. Es ist die Besitzergreifung des Verstandes, in klarer und lebendiger Weise, von einem aus scheinbar mehreren gleichzeitig möglichen Objekten oder Gedankengängen. Sein Wesensinhalt liegt in der Fokussierung und Konzentration des Bewusstseins. Dies beinhaltet die Vernachlässigung einiger Dinge, um andere Dinge effektiv abhandeln zu können“

Mit der Aussage, dass jeder wisse, was Aufmerksamkeit ist, weist James darauf hin, dass es für den Menschen eine gewohnte und natürliche Sache ist, dass sich das Bewusstsein auf bestimmte Dinge fokussiert. Dieser Umstand hilft jedoch wenig bei der Ableitung eines Modells, da der eigentliche Prozess, der die Aufmerksamkeit erzeugt, für den Menschen völlig unbewusst abläuft. Um diesen Prozess zu verstehen, bedarf es genauerer Untersuchungen. Erkenntnisse diesbezüglich ergeben sich sowohl aus der physiologischen Betrachtungen des Sehsystems, d.h. Auge, Netzhaut, Sehnerv, sowie die beteiligten Areale des Gehirns, als auch aus psychologischen Verhaltensexperimenten, bei denen die visuelle Wahrnehmung von Versuchspersonen unter festgelegten Rahmenbedingungen untersucht wird.

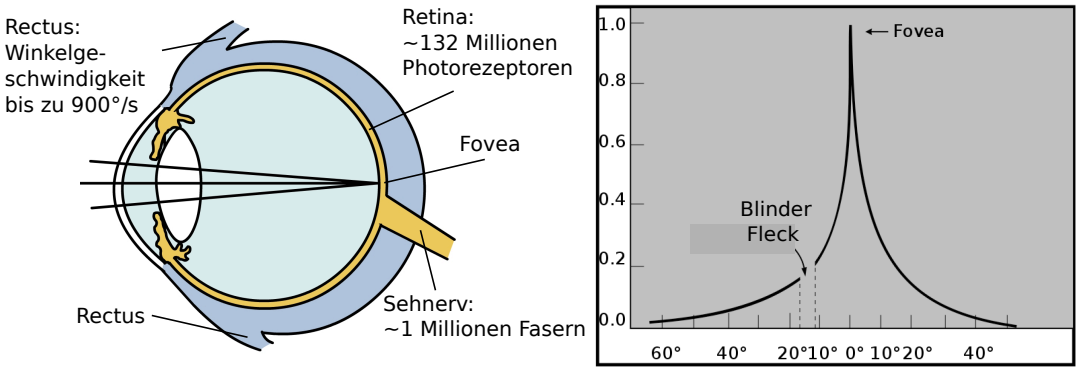

Abbildung 4.1.1: Links: Schematische Darstellung des menschlichen Auges[1]. Rechts: Verteilung der Photorezeptoren auf der Retina[2].

Bei der *selektiven Aufmerksamkeit* (Jenkin und Harris [1], S. 2) wird davon ausgegangen, dass die Aufmerksamkeit auf bestimmte Inhalte gelenkt wird, während andere Inhalte ausgeblendet werden. Eine wichtige Rolle hierbei spielt die Funktionsweise des Auges, wie in Abbildung 4.1.1 illustriert wird. Wie eingangs bereits erwähnt wurde, verfügt das Auge über mehr als 132 Millionen Photorezeptoren. Diese sind auf der Retina jedoch sehr ungleich verteilt. Die Retina hat einen zentralen Bereich mit sehr hoher Sehschärfe, die Fovea Centralis. In den umliegenden Bereichen ist die Wahrnehmung hingegen unscharf. Man unterscheidet entsprechend zwischen fovealer und peripherer Wahrnehmung (Jenkin und Harris [1], S. 87 ff.). Die Augenmuskeln (Recti) können den Augapfel mit einer Geschwindigkeit von bis zu 600 Grad pro Sekunde bewegen. Betrachtet der Mensch eine Szene, bewegen sich beide Augen parallel und fixieren in schneller, reflexartiger Abfolge mehrere Punkte im Sichtfeld. Diese Punkte werden *Fixationspunkte* und die Bewegungen *Sakkaden* (Jenkin und Harris [1], S. 24 ff.) genannt. Sakkadische Bewegungen laufen in einem Millisekundenbereich ab und geschehen für den Betrachter völlig unbewusst. Nachweisen lassen sie sich mittels hochfrequenter Videokameras. Mit Hilfe einer geeigneten Vorrichtung, die typischerweise am Kopf des Probanden angebracht und auf diesen justiert wird, lassen sich dabei auch die Blickziele ermitteln. Dies wird Eye-Tracking genannt (bspw. [59]). Der Betrachter rückt die Punkte, von denen eine visuelle Stimulation ausgeht, nacheinander in die foveale Wahrnehmung und bildet sich so unbewusst eine mentale Repräsentation der betrachteten Szene. Die Aufmerksamkeit wechselt dabei von einem Fixationspunkt zum nächsten. Ein Wechsel zwischen zwei Punkten wird entsprechend *Aufmerksamkeitsverschiebung* (Jenkin und Harris [1], S. 23 ff.) genannt. Je stimulierender ein bestimmter Punkt ist, desto schneller

1 https://commons.wikimedia.org/wiki/File:Three_Main_Layers_of_the_Eye.png, zuletzt aufgerufen am 06.09.2015.

2 https://commons.wikimedia.org/wiki/File:AcuityHumanEye.svg, zuletzt aufgerufen am 06.09.2015.

wird dieser wahrgenommen werden. Die Zahl der Punkte, die fixiert werden, ist dabei im Verhältnis zur Größe des Sichtfeldes eher gering. Die sakkadischen Bewegungen des Auges verdeutlichen bereits, dass die Aufmerksamkeit auf bestimmte Szeneninhalte gelenkt wird. Augenbewegungen und Aufmerksamkeit müssen aber nicht zwingend zusammenfallen (Jenkin und Harris [1], S. 24 ff.). Bewegen sich die Augen wie oben beschrieben zum Stimulus hin, spricht man von offener Aufmerksamkeit. Dagegen gibt es noch den Begriff der verdeckten Aufmerksamkeit (Jenkin und Harris [1], S. 22). Bei dieser wird ein Stimulus wahrgenommen, ohne dass dieser in die foveale Wahrnehmung gerückt wird. Verdeckte Aufmerksamkeit tritt typischerweise auf, wenn der Betrachter aufgrund einer Tätigkeit auf andere Punkte im Sichtbereich fixiert ist. Ein Beispiel hierfür ist das Achten auf die Straße beim Autofahren. Posner et al. [60] haben mit einem Verhaltensexperiment festgestellt, dass verdeckte Wahrnehmung in Zusammenhang mit visuellen Hinweisen zu verkürzten Reaktionszeiten führt. Bei diesen Hinweisen handelt es sich im Experiment um Pfeile, die in die Richtung zeigen, wo der Zielstimulus im peripheren Sichtbereich auftauchen wird. Der Proband fixiert dabei ein Kreuz in der Mitte eines Bildschirms und muss möglichst schnell einen Knopf drücken, sobald der Zielstimulus am Bildschirmrand auftaucht.

Eine wichtige Frage zum Verständnis der visuellen Aufmerksamkeit ist die Frage danach, welche visuellen Informationen überhaupt einen Reiz beim Betrachter auslösen. Eine wesentliche Rolle spielen dabei grundlegende visuelle Eigenschaften wie beispielsweise Intensität und Farbe. Aufmerksamkeit erzeugen solche Inhalte, die in diesen Eigenschaften eine Diskontinuität entweder örtlich zu ihrer Umgebung oder zeitlich durch eine Veränderung aufweisen. Dies wird *Pop-Out-Effekt* [62] genannt. Für die Eigenschaft hervorzustechen, wird der Begriff der *Salienz* (Jenkin und Harris [1], S. 8 ff.) verwendet. Aufmerksamkeit geht also von solchen Reizen aus, die sich von ihrem Umfeld abheben und somit *salient* sind.

Treisman und Gelade [61] haben eine Theorie aufgestellt, die sie *Merkmalsintegration* nennen. Das entsprechende Modell ist in Abbildung 4.1.2 dargestellt. Es handelt sich um ein zweistufiges Modell aus einer *präattentiven* und einer *attentiven* Phase. In der präattentive Phase wird das gesamte visuelle Feld verarbeitet. Dies läuft unbewusst, schnell, mühelos und fortwährend ab. Die attentive Phase läuft hingegen bewusst ab. Sie ist langsam, ermüdend und nur auf bestimmte Bereiche des visuellen Feldes gerichtet. Durch die zwei Stufen wird eine optimierte Nutzung der begrenzten Ressourcen ermöglicht. Die Merkmalsintegrationstheorie geht davon aus, dass in der präattentiven Phase unterschiedliche Merkmalstypen in unterschiedlichen Hirnregionen getrennt voneinander verarbeitet werden. Treisman und Gelade unterscheiden Merkmale bezüglich Farbe, Orientierung, Formung, Bewegung und Tiefe. Diese werden im abgebildeten Modell durch Merkmalskarten dargestellt, die parallel aus der Szene extrahiert und getrennt

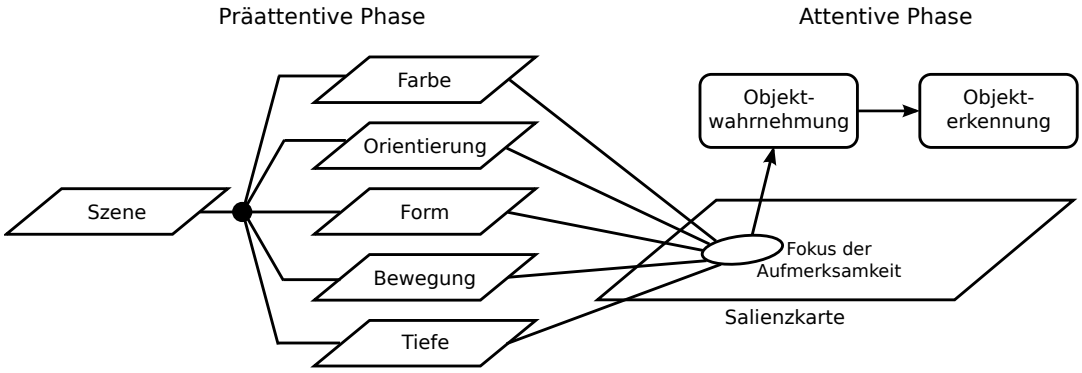

Abbildung 4.1.2: Das Merkmalsintegrationsmodell nach Treisman und Gelade [61].

auf ihre Salienz hin überprüft werden. Die Merkmalskarten werden anschließend zu einer einzigen Salienzkarte integriert. Auf dieser werden dann die salientesten Bereiche ermittelt. In der attentiven Phase werden diese Bereiche schließlich im Bewusstsein wahrgenommen und erkannt. Die getrennte Verarbeitung der unterschiedlichen Merkmalstypen lässt sich mittels einfacher Verhaltensexperimente belegen. Abbildung 4.1.3 zeigt hierfür Beispiele. Die Beispiele setzen sich aus vier verschiedenen Grundelementen zusammen, rote Rechtecke, blaue Rechtecke, rote Kreise und blaue Kreise. In jedem Beispiel sind jeweils zwei angrenzende Regionen dargestellt, die aus jeweils zwei unterschiedlichen Grundelementen bestehen. In Beispiel (a) kann die Grenze zwischen den roten und blauen Elementen präattentiv ausgemacht werden. Auch in Beispiel (b) kann eine Trennung zweier Bereiche in Kreise und Rechtecke präattentiv vorgenommen werden. In Beispiel (c) besteht die linke Region aus blauen Quadraten und roten Kreisen, die Rechte besteht aus roten Quadraten und blauen Kreisen. Obwohl das Beispiel im Grunde nicht komplexer ist als bei (a) und (b), fällt es dem Betrachter hier wesentlich schwieriger die Grenzlinie auszumachen. Der Grund hierfür ist, dass zur Trennung zwei unterschiedliche Merkmalstypen verarbeitet werden müssen. Dies kann das Sehsystem in der präattentiven Phase nicht leisten, da die Merkmalstypen getrennt voneinander verarbeitet werden. Erst in der attentiven Phase, nachdem die Merkmalstypen wieder integriert wurden, kann die Trennung der Bereiche gelingen. Aufbauend auf der Merkmalsintegrationstheorie stellen Koch und Ullman [63] eine Theorie zur Aufmerksamkeitsverschiebung auf, die sie *Winner-Take-All*-Strategie nennen. Diese besagt, dass immer der salienteste Punkt auf der Salienzkarte die gesamte Aufmerksamkeit bekommt. Dieser Punkt wird sodann auf der Salienzkarte unterdrückt. Das bedeutet, es findet eine gewisse Gewöhnung statt, die verhindert, dass Punkte mehrmals hintereinander in den Fokus der Aufmerksamkeit rücken. Dies wird als *Inhibition der Rückkehr* (Jenkin und Harris [1], S. 288 f.) bezeichnet. Bei der nächsten Auf-

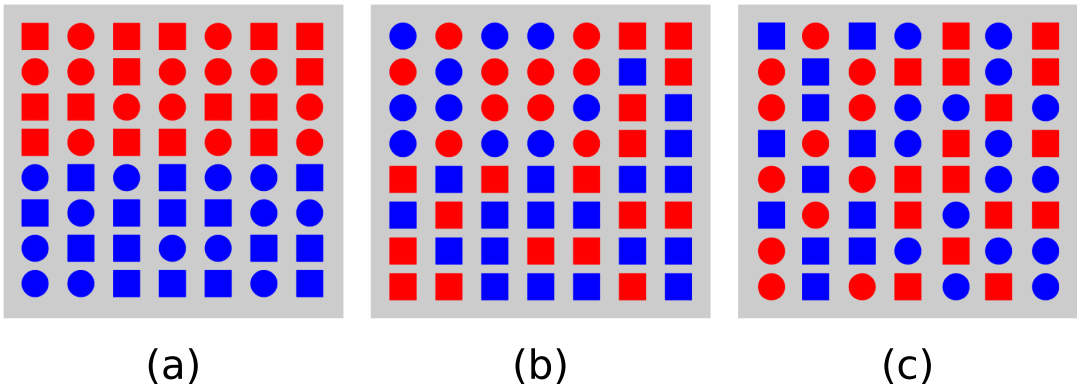

Abbildung 4.1.3: Beispielexperimente zur Merkmalsintegrationstheorie. (a) Grenze kann anhand der Farbe präattentiv ausgemacht werden. (b) Grenze kann anhand der Formung präattentiv ausgemacht werden. (c) Grenze kann nicht präattentiv ausgemacht werden.

merksamkeitsverschiebung wird dann der Punkt fixiert, der nun am salientesten ist usw.

Weitere Belege für ein zweistufiges Aufmerksamkeitsmodell finden sich bei Mishkin und Ungerleider [64]. Diese haben auf Basis von Läsionsstudien an Affen die sogenannte *Zwei-Fluss-Hypothese* aufgestellt. Dieses weitläufig anerkannte Hypothese (vgl. Jenkin und Harris [1], S. 63 ff.) stützt das zweistufige Aufmerksamkeitsmodell und besagt, dass die Lokalisierung von Objekten in einem anderen Hirnareal abläuft als die Erkennung der selben. Die Lokalisierung findet in einem weiter hinten gelegenem Pfad des Gehirns statt, der als *dorsaler Fluss* bezeichnet wird. Der weiter vorne gelegene Pfad zur Erkennung wird *ventraler Fluss* genannt. Ein weiterer Beleg für das zweistufige Modell besteht im Phänomen des *Aufmerksamkeitsblinzeln* [65]. Bei dem entsprechenden Verhaltensexperiment werden dem Probanden in gleicher, schneller Abfolge an der gleichen Stelle visuelle Reize präsentiert. Er soll dabei einen definierten Zielreiz und die auf diesen Zielreiz folgenden Reize identifizieren. Hierbei zeigte sich, dass nach Identifizierung des Zielreizes häufig ein Aufmerksamkeitsdefizit eintritt, welches mindestens eine Dauer von 270 ms aufweist. In diesem Zeitraum werden keine weiteren Reize wahrgenommen. Chun und Potter [66] leiten hieraus ab, dass nach einem ersten Prozess, der die Reize aufnimmt, ein zweiter Prozess folgt, der den jeweiligen Reiz mit mehr Tiefe verarbeitet. Dieser zweite Prozess kann mehr Zeit in Anspruch nehmen, wodurch das Aufmerksamkeitsblinzeln entsteht. Es lässt sich darüber hinaus auch experimentell zeigen, dass Aufmerksamkeit eine Voraussetzung ist, um Änderungen im Sichtfeld bewusst wahrzunehmen [67]. Dabei betrachtet der Proband eine Szene, in der ein plötzliches Ereignis eintritt. Beispielsweise kann eine Person durch das Bild laufen. Gleichzeitig verändert sich ein anderes Detail der Szene. Letzteres kann mit einer relativ hohen Wahrscheinlichkeit nicht wahrgenommen werden, da die Aufmerksamkeit

des Betrachters durch das dominierende Ereignis abgelenkt ist. Dieses Phänomen wird *Veränderungsblindheit* genannt und zeigt, dass nicht alle Veränderungen im Sichtfeld parallel verarbeitet werden können.

Bei dem oben beschriebenen Pop-Out-Effekt wird die Aufmerksamkeit ausschließlich von den durch die Photorezeptoren aufgenommenen visuellen Daten und ihren grundlegenden Eigenschaften bestimmt. Die einzelnen Stimuli konkurrieren um Aufmerksamkeit. Da es sich dabei um eine rein datengetriebene Form der Aufmerksamkeitserzeugung handelt, wird hierbei von einem *Bottom-Up*-Prozess gesprochen (Jenkin und Harris [1], S. 22, vgl. hierzu auch Abschnitt 3.2). Daneben wird von Desimone und Duncan [68] die Annahme aufgestellt, dass das Konkurrieren der Stimuli durch weitere Prozesse beeinflusst wird. Dies wird als *beeinflusste Konkurrenz-Theorie* bezeichnet. Die Beeinflussung kann durch weitere Bottom-Up- oder auch Top-Down-Prozesse entstehen. Letztere hängen dann von den Erfahrungen des Betrachters ab. Eng mit der Theorie der beeinflussten Konkurrenz verknüpft ist die Frage, ob Aufmerksamkeit von einzelnen Stimuli oder von Objekten erzeugt wird. Letzteres wird als *objektbasierte Aufmerksamkeit* (Jenkin und Harris [1], S. 9) bezeichnet. Vecera und Farah [69] haben experimentelle Belege einer objektbasierten Aufmerksamkeit gefunden. Bei ihren Verhaltensexperimenten, hat sich gezeigt, dass der Effekt der konkurrierenden Stimuli entfällt, wenn mehrere Details beobachtetet werden, die zum selben Objekt gehören. Dies ist unabhängig davon, ob sich die entsprechenden Stimuli im fovealen Sichtbereich befinden, oder weiter auseinander liegen. Duncan et al. [70] entwickelten hierzu die Theorie der *integrierten Konkurrenzhypothese*, die besagt, dass bei mehreren Objekten im Sichtfeld diese um Aufmerksamkeit konkurrieren. Es stellt sich dann die Frage, welche Objektinformationen in der präattentiven Phase ausgewertet werden. Schließlich kann ein Objekt keine Aufmerksamkeit erlangen, bevor es nicht als solches erkannt wurde. Rensink [71] stellt hierzu eine visuelle *Kohärenztheorie* auf. Diese besagt, dass in der präattentiven Phase Stimuli zu *Proto-Objekten* zusammengefasst werden, wenn diese örtlich kohärent sind. Demnach handelt es sich bei einem Proto-Objekt um eine Region im Sichtfeld des Betrachters, bei der es sich aufgrund ihrer grundlegenden visuellen Eigenschaften mit einer gewissen Wahrscheinlichkeit um ein Objekt handelt. Die Theorie stützt sich auf Verhaltensexperimente, die sich das Phänomen der Veränderungsblindheit zu Nutze machen. Hierbei stellte sich heraus, dass der Effekt der Veränderungsblindheit bei Veränderungen, die das Objekt betreffen, welches im Fokus der Aufmerksamkeit liegt, tendenziell entfällt. Eine Annahme ist, dass Gestalt-Faktoren hierbei eine wichtige Rolle spielen. Die Gestaltpsychologie [72] beschreibt Regeln, nach denen der Mensch Strukturen in seiner Wahrnehmung ausmacht. Zu diesen Faktoren zählen *Ähnlichkeit*, *Nähe* und *Geschlossenheit*. Mit ihren Verhaltensexperimenten haben Kimchi et al. [73] festgestellt, dass ein gesuchter Stimulus schneller gefunden wird, wenn dieser Teil eines nach den Gestaltgesetzen erstellten Objekts ist, als wenn kein Objekt

präsent ist. Langsamer wird er hingegen gefunden, wenn ein Objekt präsent ist und der gesuchte Stimulus nicht Teil des Objekts ist. Während die Kohärenztheorie einen reinen Bottom-Up-Prozess beschreibt, postuliert Desimone [74], dass das sogenannte *visuell arbeitende Gedächtnis*, welches im visuellen Kortex liegt, einen direkten Einfluss auf die objektbasierte Aufmerksamkeitserzeugung hat. Hiernach wird die präattentive Phase durch bereits gelernte Informationen durch ein Top-Down-Feedback beeinflusst.

Zuletzt soll noch auf das Phänomen der *Unaufmerksamkeitsblindheit* [75] eingegangen werden. Bei diesem Phänomen blendet eine Person Dinge in ihrem Sichtfeld aus, die sie normalerweise wahrnehmen würde. Dies geschieht, weil ihre Aufmerksamkeit auf andere Dinge gerichtet ist. Der Betrachter ist also abgelenkt und gegenüber bestimmten Inhalten, die er nicht erwartet, blind. Bei diesem Phänomen spielen die Voraussetzungen des Beobachters eine wichtige Rolle, d.h. mit welcher Erwartung und welcher Intention eine Szene betrachtet wird. Dies wird als *gerichtete Aufmerksamkeit* bezeichnet (Jenkin und Harris [1], S. 2). Es gibt hierzu eindrucksvolle Experimente, die zeigen, wie stark Aufgabenorientiertheit die Wahrnehmung beeinflussen kann. Ein prominentes Beispiel ist das diesbezügliche Verhaltensexperiment von Simons und Chabris [76]. Bei diesem bekommen Probanden eine Videosequenz vorgeführt, bei der sechs Personen durcheinander laufen, von denen Drei ein weißes und Drei ein schwarzes T-Shirt tragen. Die Personen im weißen T-Shirt spielen sich fortwährend einen Ball zu. Die Aufgabe des Probanden ist es, die Anzahl der Zuspiele zu zählen. Aufgrund dieser Aufgabe nehmen 75% der Probanden nicht wahr, dass irgendwann eine Person im Gorillakostüm durch die Szene läuft. Bei vorherigem Hinweis, dass etwas unerwartetes passieren könnte, fällt der Anteil auf 25%. Diese und ähnliche Experminete führen zu der Diskussion, ob Aufmerksamkeit tatsächlich in einer frühen, rein datengetriebenen Phase erzeugt wird, oder ob relevante Inhalte teilweise erst zu einer späteren Phase der Informationsverarbeitung in die bewusste Wahrnehmung vordringen (Jenkin und Harris [1], S. 197 ff.). Schließlich kann der Gorilla erst nach einer tiefergehenden Auswertung der Sehinhalte ausgeblendet werden. Spinks et al. [77] nehmen an, dass das oben bereits erwähnte visuell arbeitende Gedächtnis eine wichtige Rolle spielt und führen hierzu Messungen mittels Resonanztomographie (fMRI) an, bei denen eine Aktivität in den entsprechenden Hirnarealen aufgezeichnet wurde. Simons [78] schlägt dagegen eine andere Deutung des Phänomens der Unaufmerksamkeitsblindheit vor. Diese besagt, dass die Informationen zwar bewusst wahrgenommen, da sie aber für die Aufgabenstellung irrelevant sind, nicht im Gedächtnis abgelegt werden. Die Unaufmerksamkeitsblindheit wird somit in eine Unaufmerksamkeitsamnesie umgedeutet.

4.2 MATHEMATISCHE AUFMERKSAMKEITSMODELLE

Bei einem mathematischen Aufmerksamkeitsmodell werden Prinzipien der Aufmerksamkeit mathematisch formuliert, sodass sie in Hard- und Software umgesetzt werden können. Der Zweck eines solchen Modells kann es zum einen sein, Theorien zur menschlichen Aufmerksamkeit zu überprüfen. Ziel dabei ist es, die visuelle Wahrnehmung des Menschen besser zu verstehen. Ein mathematisches Modell kann hierzu direkt mit menschlichem Verhalten verglichen werden. Häufig ist es dabei das Ziel, menschliche Augenbewegungen vorherzusagen (bspw. [79]). Der andere Zweck für ein solches Modell, der in dieser Arbeit im Vordergrund steht, ist der Einsatz in technischen Systemen, die über Kamerasensoren verfügen. Während sich im ersten Fall die Modelle ausschließlich an der menschlichen Wahrnehmung orientieren werden, können im letzteren Fall auch rein mathematische, statistische oder signaltheoretische Erwägungen in ein Modell miteinfliessen. Die Anwendungsmöglichkeiten für technische Systeme umfassen die Implementierung intelligenter Kameras [5], Bild- und Videokompression [6], Objekt-Tracking [7], Objektsuche [8] [9] und Objekterkennung (siehe Kapitel 5.2).

Wie im letzten Abschnitt bereits erläutert wurde, besteht ein Zusammenhang zwischen Aufmerksamkeit und Salienz. Bei der Erstellung eines Aufmerksamkeitsmodells stellt sich also die Frage, wie Salienz genau definiert ist und wie man sie messen kann. Ein entsprechendes Verfahren wird im Folgenden *Salienzdetektor* genannt. Für den Entwurf eines Salienzdetektor gilt es zum einen die betrachteten Salienzmerkmale zu definieren und zum anderen festzulegen, wie diese ausgewertet werden sollen. Hierbei ist stets zu bedenken, dass der Vorteil eines Aufmerksamkeitsmodells in der zeiteffizienten Vorfilterung liegt. Bei den Salienzmerkmalen ist es deshalb wichtig, dass sie mit vergleichsweise geringem Aufwand extrahiert und verarbeitet werden können, damit die Salienzdetektion überhaupt sinnvoll zur Effizienzsteigerung komplexerer Verfahren eingesetzt werden kann. Ausgabe eines Salienzdetektors ist in der Regel eine sogenannte Salienz- oder Aktivierungskarte. Diese ordnet jeder Position des Bildes einen Salienzwert zu. Die Auflösung der Salienzkarte kann dabei auch geringer ausfallen, als die des Originalbildes. Teil eines Aufmerksamkeitsmodells ist ein Verfahren, nach dem auf Basis der Salienzkarte die salienten Punkte ausgewählt werden, denen Aufmerksamkeit zuteil wird. Bei der Auswahl spielt auch die Reihenfolge eine Rolle. Als allgemeines Prinzip gilt hierbei, dass je höher ein Salienzwert ist, desto eher wird dieser angesteuert werden. Aus einer Reihe von salienten Punkten lässt sich auf diese Weise analog zur menschlichen Aufmerksamkeit eine Sakkade von Fixationspunkten bilden.

Abhängig von der Vorgehensweise lassen sich verschiedene Aufmerksamkeitsmodelle in unterschiedliche Kategorien einteilen. Bei Borji und Itti [80] findet sich eine umfangreiche Übersicht über verschiedene Aufmerksamkeitsmodelle aus der Literatur. Die Autoren identifizieren bestimmte Eigenschaften, die für eine

Einteilung dieser verschiedene Modelle wesentlich sind. Hierzu zählt unter anderem die Unterscheidung, ob ein Modell Bewegungsinformationen einbezieht oder nicht. Da Bewegungsinformationen Veränderungen in einer Szene beschreiben, besteht eine gewisse Wahrscheinlichkeit, dass sie von Bedeutung sind. Von daher ist bei der Betrachtung von Videosequenzen die Einbeziehungen von Bewegung als Merkmalstypus sinnvoll. Eine weitere Unterscheidung ist die zwischen offener und verdeckter Aufmerksamkeit. Offene Aufmerksamkeit bedeutet, dass das technische System auf Aufmerksamkeit reagiert. Verfügt ein System über eine entsprechende Motorik, kann es saliente Inhalte in den Fokus rücken. Beim Menschen entspricht dies der Verschiebung des fovealen Sichtbereichs zwischen unterschiedlichen Fixationspunkten. Bei Modellen, die nicht aktiv reagieren, handelt es sich entsprechend um verdeckte Aufmerksamkeit. Die Unterscheidung zwischen offener und verdeckter Aufmerksamkeit ist interessant für aktive Kamerasysteme, wie sie beispielsweise bei der Robotik zu Einsatz kommen. Die Kamera kann auf einen interessanten Szeneninhalt eingestellt werden, um diesen näher zu untersuchen (bspw. [81]).

Eine für diese Arbeit wichtige Unterscheidung ist die zwischen ortsbasierten und objektbasierten Modellen. Während bei ortsbasierten Modellen Aufmerksamkeit punktuell von einzelnen Merkmalen ausgeht, geht sie bei objektbasierten Modellen entsprechend von Objekten aus, also von Bildregionen, die logische Einheiten repräsentieren. Analog zur menschlichen Wahrnehmung kann hierfür der Begriff des Proto-Objekts verwendet werden. Für das Thema der Objekterkennung ist der objektbasierte Ansatz vorteilhaft, da ein Proto-Objekt ohne weitere Arbeitsschritte direkt von einem Klassifizierer ausgewertet werden kann. Beispiele für objektbasierte Modelle werden im nächsten Kapitel in Abschnitt 5.1.3 betrachtet.

Ein weiteres wesentliches Unterscheidungsmerkmal bei Aufmerksamkeitsmodellen ist, ob diese einen Bottom-Up- oder Top-Down-Ansatz verfolgen. Während bei Bottom-Up-Ansätzen ausschließlich grundlegende Merkmale betrachtet werden, wird bei Top-Down-Ansätzen zusätzliches Vorwissen miteinbezogen. Dies kann als Analogie zur gerichteten Aufmerksamkeit bei der menschlichen Wahrnehmung aufgefasst werden. In der Regel wird das Vorwissen durch das Auswerten von Beispielbildern erworben. Es kann sich im Sinne einer gerichteten Aufmerksamkeit auf Merkmale bestimmter Ziele beziehen, die es zu detektieren gilt. Unabhängig von bestimmten Zielmerkmalen kann auch szenisches Vorwissen verwendet werden, welches Aufschluss darüber gibt, wo interessante Bildinhalte am ehesten zu finden sind. Eine weitere Möglichkeit ist es, Annahmen über die Rahmenbedingungen zu treffen, unter denen eine Bildaufnahme entstanden ist. Hierdurch können Rückschlüsse auf typische Kontrastwerte, Lichtverhältnisse oder auch Positionen und Größen bestimmter Inhalte gezogen werden. Im nächsten Kapitel werden in Abschnitt 5.1.4 Top-Down-Verfahren aus der Literatur betrachtet.

4.3 DAS MERKMALSINTEGRATIONSMODELL NACH ITTI UND KOCH

Eines der bekanntesten und einflussreichsten Aufmerksamkeitsmodelle ist das Modell von Itti et al. [57], welches sich stark an der in Abschnitt 4.1 beschriebenen Merkmalsintegrationstheorie von Treisman und Gelade [61] orientiert. Abbildung 4.3.1 zeigt eine schematische Darstellung des Modells. Aus dem Eingabebild werden parallel mehrere Merkmalskarten extrahiert, die die Salienzen der unterschiedliche Merkmalstypen wie Intensität, Farbe und Orientierung repräsentieren. Diese werden anschließend zu einer einzigen Salienzkarte integriert. Zur Extraktion der Merkmalskarten wird das Eingabebild zunächst in den Scale-Space unter Verwendung einer neunstufigen Gauß-Pyramide überführt (siehe Abschnitt 2.1). Es werden insgesamt neun Pyramiden generiert, $I(\sigma)$ für die Intensität, $R(\sigma)$, $G(\sigma)$, $B(\sigma)$, $Y(\sigma)$ für die Farbkanäle Rot, Grün, Blau und Gelb, sowie vier Gabor-Pyramiden $O(\sigma,\theta)$ für Orientierungen mit $\theta \in \{0°, 45°, 90°, 135°\}$. Zur Berechnung eines Merkmals wird die *Zentrum-Umgebungs-Hypothese* verfolgt. Diese orientiert sich an dem in Abschnitt 4.1 beschriebenen Pop-Out-Effekt. Das Zentrum ist ein Bildpunkt auf einer fein aufgelösten Stufe der Pyramide, während der örtlich korrespondierende Punkt auf einer gröber aufgelösten Stufe die Umgebung darstellt. Die Differenz aus Zentrum und Umgebung bildet das Merkmal. Mit $\ominus$ als Operator für die stufenübergreifende Differenz ergeben sich die insgesamt 42 Merkmalskarten nach den folgenden Formeln:

$$\mathcal{I}(c,s) = |I(c) \ominus I(s)| \tag{4.3.1}$$

$$\mathcal{RG}(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{4.3.2}$$

$$\mathcal{BY}(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))| \tag{4.3.3}$$

$$\mathcal{O}(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)| \tag{4.3.4}$$

Hierbei sind c und s die jeweiligen Skalierungsstufen mit $c \in \{2,3,4\}$ und $s = c + \delta, \delta \in \{3,4\}$. Zur Bildung der Salienzkarte werden die unterschiedlich skalierten Merkmalskarten für Intensität, Farbe und Orientierung jeweils zunächst durch skalierungsübergreifende Addition, $\oplus$, zu sogenannten Auffälligkeitskarten kombiniert:

$$\overline{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s)) \tag{4.3.5}$$

$$\overline{\mathcal{C}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))] \tag{4.3.6}$$

$$\overline{\mathcal{O}} = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta)) \right) \tag{4.3.7}$$

Die Auffälligkeitskarten werden im nächsten Schritt entsprechend dem Prinzip der Merkmalsintegration zu einer einzigen Salienzkarte integriert. Da hierbei

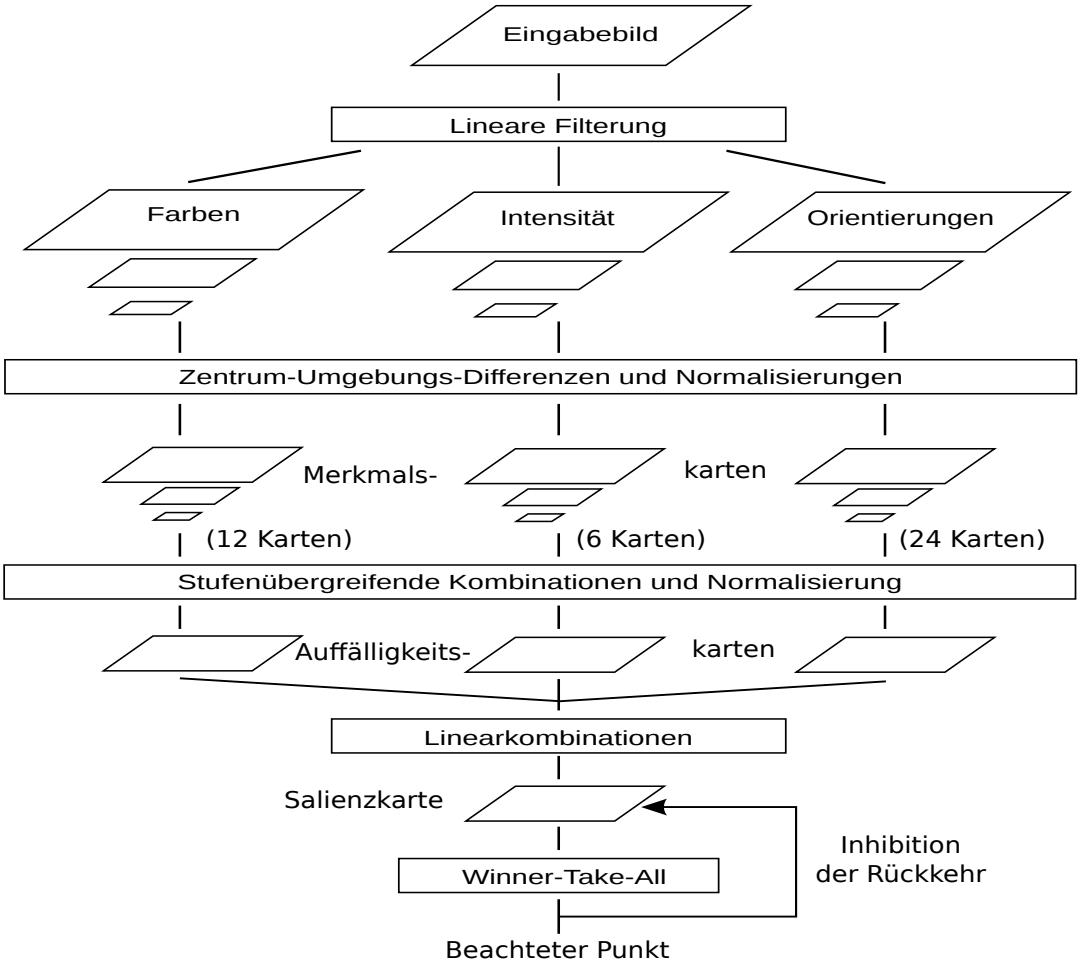

Abbildung 4.3.1: Das Merkmalsintegrationsmodell nach Itti et al. [57].

unterschiedliche Merkmalstypen kombiniert werden, muss der Wertebereich und die Werteverteilung der jeweiligen Karten berücksichtigt werden, um zu verhindern dass ein bestimmter Merkmalstyp den Prozess dominiert. Es muss also eine Möglichkeit geschaffen werden, die unterschiedlichen Merkmalstypen ausgewogen miteinander vergleichen zu können. Zu diesem Zweck wird ein Normalisierungsoperator, $\mathcal{N}(.)$, eingesetzt. Dieser wird auf jede Auffälligkeitskarte angewendet, bevor diese schließlich entsprechend folgender Formel zu einer einzigen Salienzkarte zusammengefasst werden:

$$\mathcal{S} = \frac{1}{3}\mathcal{N}(\overline{\mathcal{I}}) + \mathcal{N}(\overline{\mathcal{C}}) + \mathcal{N}(\overline{\mathcal{O}}) \tag{4.3.8}$$

Der Normalisierungsoperator überführt die Auffälligkeitskarte zunächst in einen einheitlichen Wertebereich $[0..M]$. Anschließend wird das globale Maximum M und der Durchschnitt über alle lokalen Maxima $\overline{m}$ der Karte bestimmt. Schließlich wird die Karte mit $(M - \overline{m})^2$ multipliziert. Hierdurch wird erreicht, dass nicht der absolute Wert eines lokalen Maximums ausschlaggebend für seine

Salienz ist, sondern sein Wert relativ zu den übrigen lokalen Maxima der Karte. Nachteilig hieran ist, dass für die Normalisierung der gesamten Karte nur ein Gewicht verwendet wird. Für eine Verbesserung diesbezüglich haben Itti und Koch deshalb in späteren Arbeiten [82] [15] eine alternative Vorgehensweise vorgeschlagen, bei der ein Merkmal in Relation zu seiner lokalen Umgebung gesetzt wird. Hierzu wird ein DoG-Bandpassfilter eingesetzt (siehe Abschnitt 2.1.2). Dieses hebt lokale Merkmale an und unterdrückt gleichzeitig sowohl die niederfrequente lokale Umgebung als auch sehr hochfrequente Rauscheinflüsse.

Auf Basis der Salienzkarte, $\mathcal{S}$, werden schließlich iterativ die Fixationspunkte bestimmt. Hierbei wird bei jeder Iteration die Winner-Take-All-Strategie angewendet. Diese besagt, dass der salienteste Punkt die gesamte Aufmerksamkeit bekommt. Zur Unterdrückung der bereits angesteuerten Punkte wird nach jeder Iteration eine Inhibitor-Maske an dem gewählten Fixationspunkt angelegt und von der Salienzkarte subtrahiert. Dies entspricht der Inhibition der Rückkehr bei der menschlichen Aufmerksamkeit.

5 VERWANDTE ARBEITEN

In diesem Kapitel werden thematisch verwandte Arbeiten aus der Literatur betrachtet. In Abschnitt 5.1 werden unterschiedliche Aufmerksamkeitsmodelle aufgezeigt. Abschnitt 5.2 behandelt das Thema der salienzbasierten Objekterkennung. Es werden Objekterkennungsverfahren betrachtet, die auf die ein oder andere Weise aufmerksamkeitsbasierte Methoden einsetzen. Abschnitt 5.3 geht auf verwandte Problemstellungen ein, die thematische Überschneidungen mit der Objekterkennung aufweisen. Hierbei werden solche Ansätze aufgezeigt, die aufmerksamkeitsbasierte Methoden zur Lösung dieser Probleme einsetzen.

5.1 AUFMERKSAMKEITSMODELLE

Im Folgenden werden unterschiedliche Aufmerksamkeitsmodelle aus der Literatur aufgezeigt. In Abschnitt 5.1.1 werden Ansätze zur Merkmalsintegration und in Abschnitt 5.1.2 Ansätze auf Basis einer Redundanzreduktion betrachtet. Abschnitt 5.1.3 beschäftigt sich mit objektbasierten Modellen. In Abschnitt 5.1.4 werden verschiedene Top-Down-Strategien erläutert. Die Einteilung der Abschnitte ist dabei nur als grobe Richtschnur zu verstehen, da es bei vielen Modellen zu thematischen Überschneidungen kommt.

5.1.1 *Merkmalsintegration*

Das in Abschnitt 4.3 vorgestellte Merkmalsintegrationsmodell von Itti et al. [57] hatte einen großen Einfluss auf viele spätere Arbeiten. Zu den einflussreichen Konzepten zählen die Merkmalsintegration, die Zentrum-Umgebungs-Hypothese sowie die Winner-Take-All-Strategie. Spätere Arbeiten beschäftigen sich häufig damit, bestimmte Aspekte dieser Konzepte zu erweitern oder zu verbessern. Diese Verfahren bauen meist auf dem in Abbildung 5.1.1 dargestellten Grundkonzept auf. Es handelt sich hier im Wesentlichen um eine generalisierte Darstellung des Merkmalsintegrationsmodells nach Treisman und Gelade [61] (siehe Abschnitt 4.1).

Ein Beispiel hierzu findet sich bei Harel et al. [83], die einen graphenbasierten Ansatz vorschlagen, bei dem jeder Punkt einer Merkmalskarte einen Knoten darstellt. Die Knoten sind jeweils mit allen anderen Knoten in beiden Richtungen verbunden. Die Kantengewichte ergeben sich aus einer Vorschrift, die die örtliche und visuelle Distanz der jeweiligen Merkmale auswertet. Durch eine Normalisierung der ausgehenden Kantengewichte zu einer Summe von Eins wird

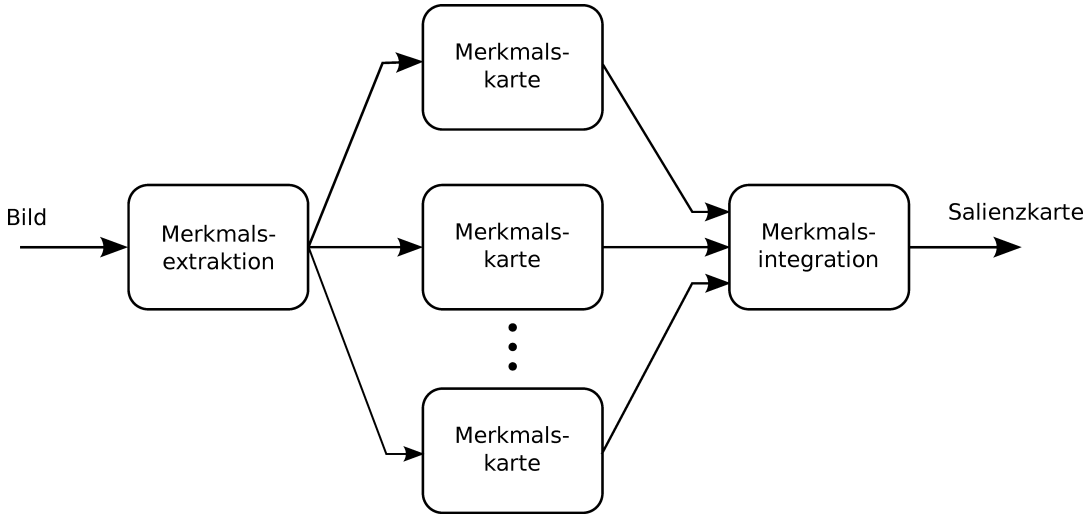

Abbildung 5.1.1: Generalisiertes Schema von Modellen auf Basis der Merkmalsintegration.

ein Markov'scher Graph (Bishop [12], S. 359 ff.) erstellt, bei dem die Knoten die Zustände und die Kantengewichte die Übergangswahrscheinlichkeiten angeben. Aus diesem Graph wird dann eine Massenverteilung ermittelt, indem wiederholt für allen Knoten zunächst die ankommende Masse über alle Eingangskanten ermittelt und diese über die ausgehenden Kanten entsprechend der Übergangswahrscheinlichkeiten gewichtet an die benachbarten Knoten verteilt wird. Ein Bereich, der sich von seiner Umgebung stark unterscheidet, wird auf diese Weise eine große Masse anziehen. Aus den Massenverteilungen der Merkmalskarten ergibt sich dann schließlich die Salienzkarte.

Ein weiteres Beispiel findet sich bei Le Meur et al. [79]. Bei diesem Verfahren wird das Bildsignal im Rahmen der Merkmalsextraktion zunächst in einen achromatischen und zwei chromatische Kanäle aufgeteilt. Auf den Kanälen wird jeweils eine Funktion zur Kontrastsensitivität angewendet. Dabei wird angenommen, dass die Amplitude einer Frequenz einen Schwellwert überschreiten muss, um wahrgenommen zu werden. Es erfolgt sodann eine Aufteilung der Kanäle in jeweils mehrere Ortsfrequenzbänder. Auf diesen werden anschließend Maskierungseffekte der menschlichen Wahrnehmung modelliert. Dabei wird davon ausgegangen, dass benachbarte Frequenzen sich gegenseitig beeinflussen. Wenn das maskierende Signal unterhalb eines bestimmten Schwellwertes liegt, kann das maskierte Signal leichter wahrgenommen werden. Als nächstes werden die achromatischen Teilbänder an jeweils den Stellen angehoben, an denen die chromatischen Teilbänder einen hohen Farbkontrast aufweisen. Zur Modellierung der Zentrum-Umgebungs-Hypothese werden anschließend DoG-Filter eingesetzt. Ein weiteres Filter wird angewendet, um solche Stellen anzuheben, die ähnlich orientierte Kanten in ihrer Umgebung aufweisen. Hierdurch soll die Eigenschaft der menschlichen Wahrnehmung modelliert werden, ähnliche Elemente zu gruppieren.

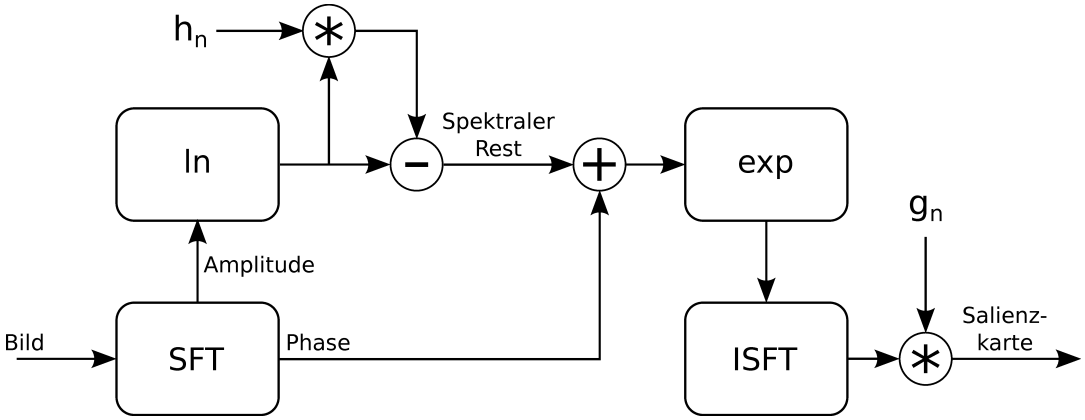

Abbildung 5.1.2: Salienz auf Basis des spektralen Rests nach Hou und Zhang [84].

Die Integration zu einer Salienzkarte erfolgt schließlich durch die Aufsummierung der achromatischen Teilbänder.

Als letztes Beispiel sei das Modell von Gao et al. [85] betrachtet. Interessant an diesem Modell ist die Einbeziehung von Bewegungsinformationen. Es werden zusätzlich zu den Merkmalskarten, die bereits bei dem Modell von Itti et al. [57] betrachtet werden, weitere Karten extrahiert. Dazu werden dreidimensionale Gabor-Filter eingesetzt, bei denen die zusätzliche Dimension die Zeitachse darstellt. Es werden Filtermasken für vier örtliche Orientierungen mit jeweils drei zeitlichen Frequenzen eingesetzt. Insgesamt werden also zwölf zusätzliche Merkmalskarten extrahiert.

5.1.2 *Redundanzreduktion*

Bei verschiedenen Modellen wird davon ausgegangen, dass die salienten Anteile eines Bildsignals eine geringe Redundanz aufweisen und somit gewisse Alleinstellungsmerkmale aufweisen, durch die sie hervorstechen. Redundante Informationen sind hingegen uninteressant. Danach lässt sich eine Salienzkarte durch die Reduktion von Redundanzen ermitteln.

Das Problem lässt sich beispielsweise auf Basis einer Histogrammdarstellung adressieren. Ein entsprechendes Modell findet sich bei Zhai und Shah [87]. Zur Berechnung der Salienzkarte wird jedem Farbwert, $I(m) \in [0,255]$, ein fester Salienzwert zugeordnet. Dies geschieht, indem zunächst das Histogramm des Bildes über die Farbwerte gebildet wird. Die Salienz bestimmt sich dann entsprechend folgender Formel:

$$S(m) = S(I(m)) = \sum_{n=0}^{255} f_n |m - n| \tag{5.1.1}$$

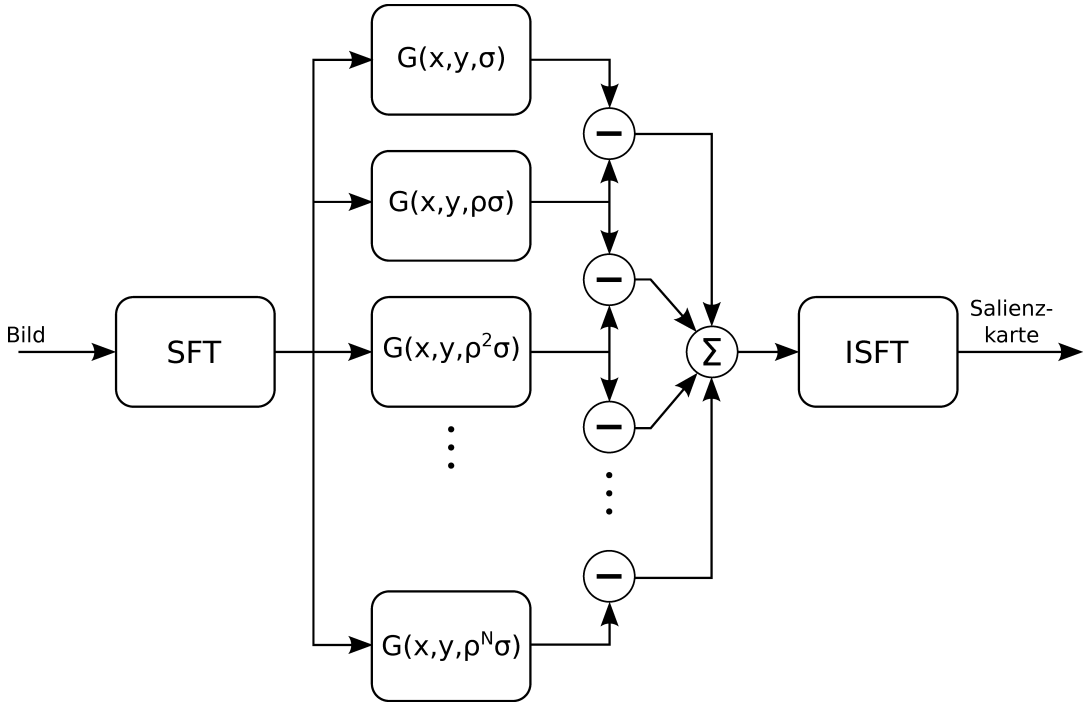

Abbildung 5.1.3: Kombination mehrerer DoG-Filter nach Achanta et al. [86].

Hierbei bezeichnet f_n den n-ten Kanal des Histogramms. Für alle Farbwerte wird also jeweils der Kontrast zum Vergleichswert mit der Häufigkeit des Vergleichswertes multipliziert. Auf diese Weise werden den redundanten Farbwerten niedrige Salienzwerte zugeordnet. Hingegen weisen Bildpunkte, die sich aufgrund ihres Farbwertes stärker vom Rest des Bildes unterscheiden, entsprechend hohe Salienzwerte auf.

Bei dem Modell von Hou und Zhang [84] wird das Problem der Redundanzreduktion im Ortsfrequenzbereich adressiert. Dass bedeutet, es werden Redundanzen aus dem Spektrum des Eingabebildes entfernt. Der verbliebene spektrale Rest stellt die interessante Innovation des Bildes dar. Zurück in den Ortsbereich transformiert, ergibt sich hieraus die Salienzkarte. Bei der Bestimmung des spektralen Rests wird ausgenutzt, dass logarithmische Spektren natürlicher Bilder im Durchschnitt einen annähernd linearen Verlauf aufweisen. Die linearen Anteile lassen sich näherungsweise durch den Einsatz eines lokalen Mittelwertfilters, h_n, extrahieren und können dann vom Spektrum subtrahiert werden. Das Vorgehen ist in Abbildung 5.1.2 dargestellt. SFT bezeichnet hier die schnelle Fourier-Transformation und ISFT entsprechend die inverse Transformation (Gonzalez und Woods [11], S. 208 ff.). Wie zu sehen ist, wird das beschriebene Vorgehen nur auf den Amplitudengang angewendet, damit die Phasen erhalten bleiben. Nach der Rücktransformation wird die Salienzkarte schließlich noch mit einem Gauß-Filter, g_n, geglättet, um hochfrequente Störeinflüsse zu unterdrücken.

Ein Nachteil des Verfahrens ist, dass auch die hochfrequenten Signalanteile zu einem gewissen Grad unterdrückt werden. Dies führt in der Salienzkarte zu einer Unschärfe an den Rändern salienter Bereiche. Achanta et al. [86] schlagen deshalb eine alternative Vorgehensweise vor. Diese basiert auf einer Parallelschaltung von mehreren DoG-Filtern (siehe Abschnitt 2.1.2) mit angrenzenden Standardabweichungen, wie sie in Abbildung 5.1.3 dargestellt ist. Bei diesem Vorgehen wird der redundante Gleichanteil unterdrückt. Gleichzeitig bleiben die hohen Frequenzanteile durch die DoG-Bandpässe erhalten. Ein weiterer Vorteil ist, dass sich das dargestellte Schema entsprechen der folgenden Gleichung stark vereinfachen lässt:

$$\sum_{n=0}^{N-1} \Big(G(x,y,\rho^{n+1}\sigma) - G(x,y,\rho^{n}\sigma) \Big) = G(x,y,\rho^{N}\sigma) - G(x,y,\sigma). \quad (5.1.2)$$

Mit ρ^N gegen unendlich isoliert man schließlich den Gleichanteil, sodass man zu folgender, einfachen Gleichung kommt:

$$\mathcal{S}(x,y) = ||I_\mu - I_{\omega_{hc}}(x,y)||. \quad (5.1.3)$$

I_μ ist hier der Gleichanteil, d.h. der arithmetische Mittelwert des Bildes I, und $I_{\omega_{hc}}$ eine gaußgefilterte Version von I. Durch die Gauß-Filterung sollen hochfrequente Rauscheinflüsse unterdrückt werden. Dabei stellt ω_{hc} die entsprechende Grenzfrequenz dar. Mit $||\,||$ wird die L_2-Norm bezeichnet. Bei Farbbildern lässt sich die Differenz als euklidische Distanz im Farbraum bestimmen.

Guo und Zhang [6] schlagen ein Modell vor, welches Bewegungsinformationen miteinbezieht. Neben der Intensität und zwei Farbkanälen werden zeitliche Differenzbilder ausgewertet. Hierzu wird statt der herkömmlichen Fourier-Transformation eine Quaternion-Fourier-Transformation eingesetzt. Während der Fourier-Raum in der herkömmlichen Variante aus einem Real- und einem Imaginärteil besteht, setzt er sich bei der Quaternion-Variante aus einem Real- und drei Imaginärteilen zusammen. Auf diese Weise können vier Komponenten zum Zeitpunkt t entsprechend der folgenden Gleichung betrachtet werden:

$$I(t) = \frac{r(t) + g(t) + b(t)}{3} \quad (5.1.4)$$

$$RG(t) = R(t) - G(t) \quad (5.1.5)$$

$$BY(t) = B(t) - Y(t) \quad (5.1.6)$$

$$M(t) = |I(t) - I(t-\tau)| \quad (5.1.7)$$

Die Farbkomponenten, R, G, B und Y, werden entsprechend dem Modell von Itti et al. [57] ermittelt (siehe Abschnitt 4.3). Die Bewegungskomponente, M, wird durch das Differenzbild aus dem aktuellen Bild und einem vorangegangenen Bild ermittelt.

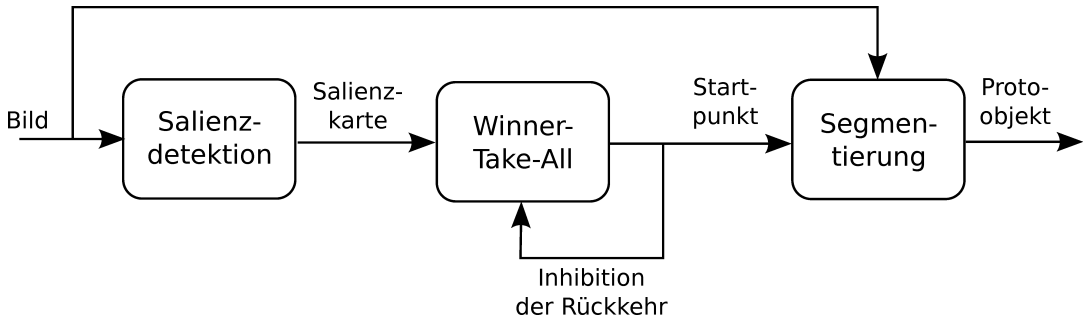

Abbildung 5.1.4: Extraktion der Proto-Objekte auf Basis der Winner-Take-All-Strategie.

Eine weitere Variante findet sich bei Li et al. [88]. Auch dort werden Redundanzen im Ortsfrequenzbereich entfernt. Es wird jedoch ein hierarchischer Ansatz verfolgt, bei dem unterschiedlich stark geglättete Versionen des Spektrums betrachtet werden. Hierdurch entstehen mehrere Salienzkarten, aus denen die Karte mit der niedrigsten Entropie ausgewählt wird. Die Idee dahinter ist, dass bei einer niedrigen Entropie viele gewöhnliche Inhalte unterdrückt wurden und nur noch die besonderen Inhalte repräsentiert werden.

5.1.3 *Objektbasierte Modelle*

Objektbasierte Modelle zeichnen sich dadurch aus, dass sie statt einer Salienzkarte eine Liste von Objektkandidaten ausgeben. Diese werden als Objekthypothesen, Proto-Objekte oder einfach als Regions-Of-Interest (ROIs) bezeichnet. Typischerweise werden sie durch Regionen oder Rechtecke angegeben. Zusätzlich kann jeder Objekthypothese ein Salienz- oder Zuversichtswert zugeordnet werden, um eine Reihenfolge unter diesen festzulegen. Zur Bestimmung von Objekthypothesen kommen unterschiedliche Strategien in Betracht, wie im Folgenden anhand von Beispielen aus der Literatur erläutert wird.

Als Erstes wird hierzu der Ansatz in Abbildung 5.1.4 betrachtet. Das Vorgehen basiert auf der Winner-Take-All-Strategie. Durch den Einsatz eines lokalen Segmentierungsverfahrens kann ein ortsbasiertes Salienzmodell erweitert werden, um Proto-Objekte zu bestimmen, indem das Proto-Objekt am jeweils salientesten Punkt heraussegmentiert wird. Entsprechende Ansätze finden sich bei Walther et al. [89] sowie Rutishauser et al. [90]. Beide Ansätze bauen auf das Modell von Itti et al. [57] auf und verwenden zur Segmentierung eine Region-Growing-Strategie. In einer späteren Arbeit verwenden Walther und Koch [91] anstelle des Region-Growing-Ansatzes ein künstliches neuronales Netz. Die Idee dahinter ist, dass sich die Segmentierung aufgrund der neuronale Netzwerkstruktur stärker an der menschlichen Wahrnehmung orientiert.

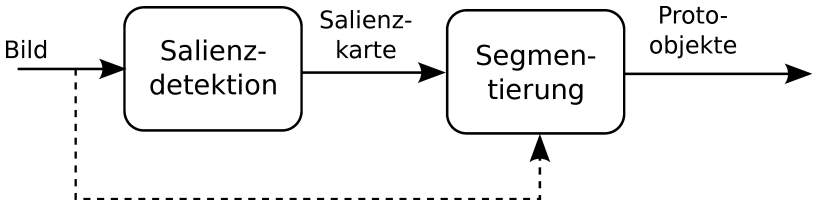

Abbildung 5.1.5: Extraktion der Proto-Objekte durch Segmentierung auf der Salienzkarte.

Abbildung 5.1.5 zeigt eine Variante, bei der die Segmentierung der Proto-Objekte nicht auf dem Eingabebild, sondern auf der Salienzkarte erfolgt. Möglich ist auch eine Kombination aus der Salienzkarte und weiteren Komponenten des Eingabebildes, was durch den gestrichelten Pfeil angedeutet wird. Ein Ansatz dieser Art findet sich bei Zhai und Shah [87]. Bei diesem wird ausgehend von einem salienten Punkt ein Rechteck ausgedehnt, um das zugehörige Objekt auf der Salienzkarte einzurahmen. Ein anderes Beispiel findet sich bei Hou und Zhang [84]. Hier wird auf der Salienzkarte eine Schwellwertsegmentierung entsprechend folgender Gleichung angewendet:

$$\mathcal{O}(x) = \begin{cases} 1, \text{wenn } S(x) > E(S(x)) \times 3 \\ 0, \text{sonst.} \end{cases} \tag{5.1.8}$$

Der Schwellwert hängt also vom Erwartungswert der Salienzwerte ab. Aus der Objektkarte, $\mathcal{O}$, können die Proto-Objekte nun als die zusammenhängenden Bereiche extrahiert werden. Bei Achanta et al. [86] wird ein komplexeres Segmentierungsverfahren auf Basis des Mean-Shift-Ansatzes [92] verwendet. Dabei werden neben der Salienzkarte noch die Intensität und zwei Farbkanäle des Eingabebildes betrachtet.

Eine andere Strategie findet sich bei Jo et al. [93]. Bei ihrem Ansatz erfolgt die Segmentierung von Proto-Objekten auf Basis von Gestalt-Faktoren [72]. Die Gestaltpsychologie beschreibt Regeln, nach denen der Mensch Strukturen in seiner Wahrnehmung ausmacht. Um Proto-Objekte zu gruppieren werden Gestalt-Faktoren für „Ähnlichkeit“, „gute Fortsetzung“ und „Nähe“ definiert. Die Idee des Ansatzes ist es folglich, die Detektion der Proto-Objekte stärker an der menschlichen Wahrnehmung auszurichten. Ein weiteres Beispiel zur gestaltbasierten Aufmerksamkeit findet sich bei Wu und Zhang [94]. Bei ihrem Ansatz werden Methoden der Gestaltpsychologie angewendet, um zusammenhängende Hintergrundbereiche auszublenden und so die Proto-Objekte hervorzuheben.

Abbildung 5.1.6 zeigt eine weitere Strategie zur Detektion von Proto-Objekten. Es handelt sich hierbei um eine vereinfachte Darstellung des Verfahrens nach Alexe et al. [95]. Zunächst wird für eine flächendeckende, sehr große Anzahl an Fensterausschnitten die Salienz bestimmt. Anschließend wird mittels eines Suchmusters eine festgelegte Anzahl an Ausschnitten durchlaufen. Hierbei werden hohe Salienzwerte und eine ausgewogene Verteilung über die Bildfläche

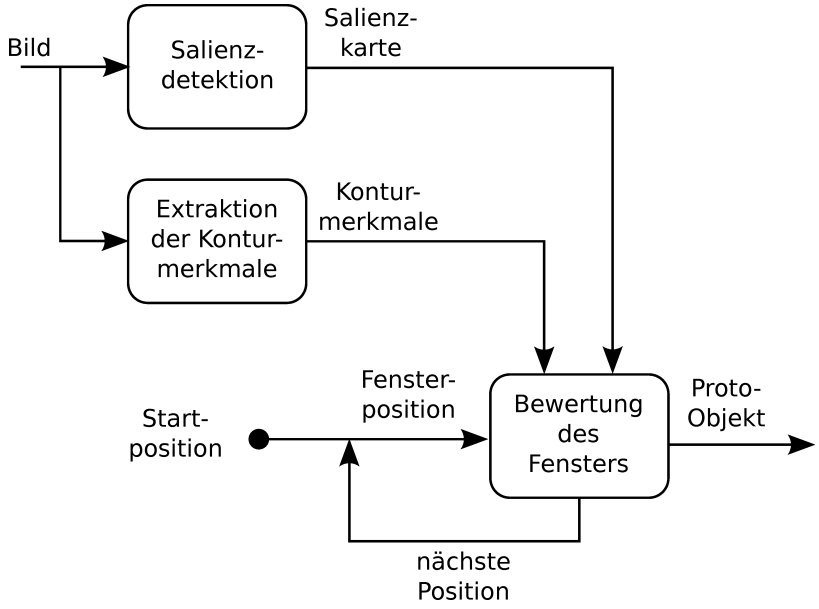

Abbildung 5.1.6: Extraktion der Proto-Objekte mittels Suchfenster. Vereinfachte Darstellung nach Alexe et al. [95].

als Kriterien angewendet. An jeder Fensterposition werden weitere Merkmale bezüglich der Kontur ausgewertet, um zu bewerten, ob es sich bei dem Inhalt des Fensters um ein Objekt handeln könnte.

Eine weitere Strategie ist in Abbildung 5.1.7 zu sehen. Bei dieser wird im ersten Schritt das Eingabebild segmentiert. Im Anschluss werden die resultierenden Regionen auf ihre Salienz hin überprüft. Dabei wird auf Basis von regionalen Salienzmerkmalen jeder Region ein Salienzwert zugeordnet. Dieses Prinzip wird bei dem von Cheng et al. [96] vorgeschlagenen Modell angewendet. Bei diesem wird das Eingabebild zunächst in disjunkte Regionen unterteilt. Der Salienzwert einer Region bemisst sich anschließend danach, wie stark sie sich von den übrigen Regionen unterscheidet. Um dies zu bemessen, wird ein visuelles Distanzmaß definiert. Für dieses wird der Farbraum quantisiert und zu jeder Region wird ein Farbhistogramm erstellt. Die visuelle Distanz zwischen zwei Regionen, r_1 und r_2, bemisst sich dann wie folgt:

$$D_r(r_1, r_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f_1(c_1(i)) f_2(c_2(j)) D(c_1(i), c_2(j)) \tag{5.1.9}$$

Dabei ist $c_k(i)$ der Farbwert des i-ten Bildpunktes der Region r_k und $f_k(n)$ ist die Häufigkeit des n-ten Farbwertes des Histogramms der Region r_k. Des Weiteren gibt die Funktion D die Distanz im Farbraum zweier Farbwerte an. Die visuelle Distanz ergibt sich folglich aus der Summe der Distanzen im Farbraum

Abbildung 5.1.7: Bestimmung der Proto-Objekte mittels Segmentierung auf dem Eingabebild und Extraktion regionaler Salienzmerkmale.

aller Bildpunkte der einen Region zu allen Bildpunkten der anderen Region. Die Salienz einer Region, r_k, ergibt sich dann zu

$$S(r_k) = \sum_{r_k \neq r_i} w_s(r_i) w_d(r_k, r_i) D_r(r_k, r_i), \tag{5.1.10}$$

wobei $w_s(r_i)$ die Größe der jeweiligen Vergleichsregion, r_i, angibt, um Unterschiede zu größeren Regionen entsprechend höher zu gewichten. Des Weiteren berücksichtigt $w_d(r_k, r_i)$ die örtliche Distanz $D_s(r_k, r_i)$ zwischen r_k und r_i, um Unterschiede zu näher liegenden Regionen stärker zu bewerten:

$$w_d(r_k, r_i) = \exp\left(\frac{-D_s(r_k, r_i)}{\sigma_s^2}\right) \tag{5.1.11}$$

Die Salienz einer Region ergibt sich also aus den visuellen Distanzen einer Region zu allen anderen Regionen in ihrer Umgebung. Dies entspricht der Zentrum-Umgebungs-Hypothese.

5.1.4 *Top-Down-Modelle*

Wie in Abschnitt 4.2 bereits erläutert wurde, kann bei Aufmerksamkeitsmodellen zwischen Bottom-Up- und Top-Down-Modellen unterschieden werden. Unter Top-Down-Modellen werden hier solche Modelle verstanden, die Annahmen treffen, die sich nicht ohne Weiteres auf natürliche Bilder im Allgemeinen anwenden lassen. Insbesondere sind Modelle gemeint, die Beispielbilder einsetzen, um Modellwissen zu beziehen. Im letzten Kapitel wurde in Abschnitt 4.1 in diesem Zusammenhang bereits auf die Annahme eingegangen, dass Aufmerksamkeit aufgabenorientiert ist, was als gerichtete Aufmerksamkeit bezeichnet wurde. Des Weiteren wurde das visuell arbeitende Gedächtnis und die Annahme angesprochen, dass es einen Einfluss auf die präattentive Phase hat. Bei entsprechenden Modellen kommt es häufig zu thematischen Überschneidungen mit der Objekterkennung, was insbesondere dann der Fall ist, wenn zur Bildung des visuellen Gedächtnisses Eigenschaften von Objekten angelernt werden. Bei einem Top-Down-Modell muss zum einen festgelegt werden, welches Modellwissen verwendet werden soll und zum anderen, wie dieses die Erzeugung der Aufmerksamkeit beeinflusst. Hierzu gibt es unterschiedliche Strategien, von denen einige im Folgenden anhand von Beispielen aus der Literatur aufgezeigt werden.

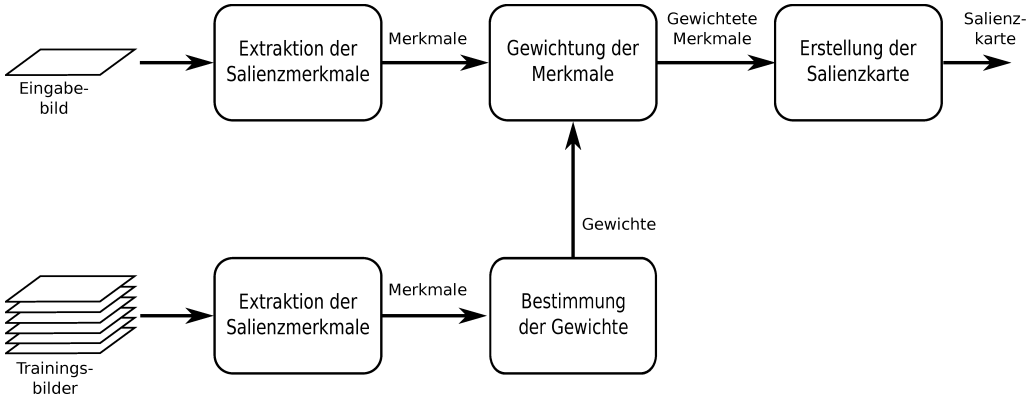

Abbildung 5.1.8: Gewichtung der Salienzmerkmale auf Basis von Trainingsbeispielen.

Eine Strategie kann es sein, Modellwissen aus Beispielbildern für bestimmte Zielobjekte zu beziehen. Beispielsweise kann im Straßenverkehr damit gerechnet werden, Fahrzeuge, Fahrradfahrer oder Fußgänger vorzufinden. Für eine entsprechende Aufgabe kann dann Modellwissen genutzt werden, um bestimmte Salienzmerkmale stärker oder schwächer zu gewichten. Abbildung 5.1.8 illustriert das Vorgehen. Ein Beispiel hierfür ist das Verfahren von Navalpakkam und Itti [97]. Dieses basiert auf das in Abschnitt 4.3 beschriebene Merkmalsintegrationsmodell nach Itti et al. [57]. Bevor die Merkmalskarten jedoch zu einer Salienzkarte integriert werden, werden diese jeweils mit unterschiedlichen Gewichten multipliziert. Die Gewichte werden zuvor mit Hilfe von Beispielbildern angelernt, wobei die Ausprägung der Objektmerkmale im Verhältnis zu den Hintergrundmerkmalen betrachtet wird. Ein ganz ähnliches Vorgehen findet sich bei Zhao und Koch [98]. Hier werden jedoch keine speziellen Zielobjekte betrachtet. Stattdessen werden die Gewichte auf Basis von Eye-Tracking-Daten angelernt. Das Ziel dabei ist es, menschliche Augenbewegungen vorherzusagen bzw. zu simulieren.

Eine andere Strategie ist es, Annahmen über bekannte Rahmenbedingungen bei der Bilderfassung zu treffen. Andreopoulos und Tsotsos [99] untersuchen hierzu die Einflüsse verschiedener Faktoren auf Keypoint-, Salienz- und Mustererkennungsverfahren. Die Betrachtung umfasst zum einen Faktoren der Aufnahmetechnik, und zwar unterschiedliche Sensoren (CCD und CMOS), Belichtungszeiten und optische Verstärkungen. Zum anderen werden Einflüsse einer variierenden Belichtung und wechselnder Perspektiven untersucht, wie sie typischerweise in einer weitgehend unkontrollierten Umgebung auftreten können. Die dort präsentierten Experimente zeigen, dass die verschiedenen Faktoren einen signifikanten Einfluss auf die Ergebnisse haben. Bei Fotografien muss jedoch nicht unbedingt von einer vollständig unkontrollierten Umgebung ausgegangen werden. Häufig wird ein Fotograf seine Aufnahmen bei guten Lichtverhältnissen machen und das Objektiv annähernd horizontal ausrichten. Des Weiteren wird

er bestimmte Inhalte, die er interessant findet, in den Fokus der Aufnahme rücken. Solche Faktoren lassen sich im Rahmen eines Top-Down-Modells direkt oder indirekt durch die Verwendung von Beispielbildern ausnutzen. Eine übliche Strategie diesbezüglich ist es, vorab Annahmen über die wahrscheinliche Lage von Bildbereichen mit interessanten Inhalten zu treffen. Wird im einfachsten Fall davon ausgegangen, dass der Fotograf interessante Inhalte in den Bildmittelpunkt rückt, kann eine Gauß-Maske auf die Salienzkarte angewendet werden, um Punkte nahe der Bildmitte anzuheben (bspw. [100] [101]). Eine andere Variante findet sich bei Alexe et al. [95]. Hier wird mittels Beispielbildern eine Statistik über die Position und Größe von Objekten erzeugt.

Bei Oliva et al. [102] wird ein Ansatz vorgeschlagen, bei dem die Lage interessanter Inhalte im Bild durch die Auswertung des szenischen Kontexts vorhergesagt wird. Mit Hilfe von Beispielbildern werden bedingte Wahrscheinlichkeiten für das auftreten lokaler Merkmale in einer Szene und das Vorfinden von Objekten an bestimmten Positionen im Bild ermittelt. Mit $\mathbf{v}_l$ als Merkmalsvektor, der einen lokalen Punkt $\mathbf{x}$ einer Szene beschreibt und $\mathbf{v}_c$ als Merkmalsvektor, der die Szene als Ganzes repräsentiert, wird die Wahrscheinlichkeitsdichtefunktion, $p(\mathbf{v}_l|\mathbf{v}_c)$, aufgestellt. Salienz wird nun als Kehrwert dieser Funktion definiert:

$$S(\mathbf{x}) = (p(\mathbf{v}_l|\mathbf{v}_c))^{-1} \tag{5.1.12}$$

Dadurch werden unwahrscheinliche Ereignisse als salient eingestuft. Diese Definition wird nun um einen Term erweitert, um kontextuelles Vorwissen über Objektpositionen einfließen zu lassen:

$$S_c(\mathbf{x}) = S(\mathbf{x})p(o, \mathbf{x}|\mathbf{v}_c) \tag{5.1.13}$$

Der Term beschreibt die Wahrscheinlichkeit, dass sich bei gegebener Szene, $\mathbf{v}_c$, ein Objekt, o, an einem bestimmten Punkt, $\mathbf{x}$, befindet. Durch Multiplikation dieses Terms mit der Salienzkarte, S, ergibt sich so die kontextuelle Salienzkarte, S_c.

Bei der letzten Top-Down-Strategie, die hier aufgezeigt werden soll, wird davon ausgegangen, dass sich das menschliche Sehsystem im Laufe seiner evolutionären Entwicklung auf die Erkennung bestimmte Muster besonders stark spezialisiert hat, sodass sie bereits in einer frühen Phase der Wahrnehmung detektiert werden können. Hieraus ergibt sich der Ansatz, entsprechende Muster im Rahmen eines Aufmerksamkeitsmodells zu betrachten. Judd et al. [100] verfolgen diese Strategie, um menschliche Augenbewegungen besser vorhersagen zu können. Sie verwenden Merkmale auf drei unterschiedlichen Komplexitätsstufen. Auf der niedrigsten Stufen werden im Rahmen einer Bottom-Up-Salienzdetektion einfache Merkmale bezüglich Intensität, Farbe und Orientierung betrachtet. Auf der mittleren Stufe wird ein Liniendetektor eingesetzt. In der Natur richten sich Objekte häufig vertikal zur Schwerkraft aus, wodurch es zu entsprechenden Linienverläufen kommt. Bei der Verwendung eines Liniendetektors wird davon

ausgegangen, dass sich das menschliche Sehsystem daran angepasst hat. Auf der höchsten Komplexitätsstufe wird ein Gesichtsdetektor und ein Personendetektor eingesetzt. Hierbei wird die Annahme getroffen, dass der Mensch im Laufe der Evolution gelernt hat, mit seinesgleichen effizient zu interagieren, und die hierfür relevanten Muster effizient erkennen kann.

5.2 AUFMERKSAMKEITSBASIERTE OBJEKTERKENNUNG

Im Folgenden werden Verfahren aus der Literatur betrachtet, die das Thema der Objekterkennung mit dem Thema der visuellen Aufmerksamkeit verknüpfen.

Elazary und Itti [3] gehen der grundlegenden Fragestellung nach, ob es sinnvoll ist, Aufmerksamkeitsmodelle und Methoden der Objekterkennung zu kombinieren. Sie stellen hierzu die These auf, dass die interessanten Objekte einer Szene salient sind (vgl. Abschnitt 1.2). Um dies zu belegen, wenden sie das Modell von Itti et al. [57] auf 24.863 Bilder der *LabelMe*-Datenbank [10] an. Die Datenbank wird deshalb als geeignet angesehen, da die Bearbeiter, die die Annotationen erstellt haben, keine konkrete Aufgabenstellung erhalten haben. Sie wurden lediglich gebeten, „nette Labels“ zu erzeugen. Die Autoren gehen deshalb davon aus, dass zumindest zum überwiegenden Teil solche Objekte annotiert wurden, die der jeweilige Bearbeiter als interessant empfunden hat. Bei der Evaluierung der Ergebnisse wird überprüft, ob die salienten Punkte, die detektiert wurden, auf einem annotierten Objekt liegen. Dies wird mit den Ergebnissen von zufällig gewählten Punkten verglichen. Es wird dabei der verzerrende Faktor berücksichtigt, dass die Tendenz besteht, eher Objekte zu annotieren, die zentral im Bild liegen. Um diesen Faktor auszugleichen, wird die örtliche Verteilung der segmentierten Objekte bestimmt und eine entsprechende Gewichtung der Salienzen vorgenommen. Die Ergebnisse zeigen, dass bei 43% der Bilder der salienteste Punkt des Bildes auf einer annotierten Region liegt. Bei 76% der Bilder ist es mindestens einer der drei salientesten Punkte. Im Vergleich erzielen zufällig gewählte Punkte lediglich eine Trefferquote von 21% bzw. 43%. Insgesamt zeigen die Ergebnisse einen signifikanten Zusammenhang zwischen interessanten Objekten und salienten Bildinhalten. Bei der Bewertung der Ergebnisse ist zu beachten, das die Größe eines Objektes nicht bestimmt wird. Für ein positives Ergebnis ist es ausreichend, wenn ein salienter Punkt irgendwo auf dem Objekt liegt.

Bei verschiedenen Ansätzen wird ein zweistufiges System aus einem Aufmerksamkeitsmodell und einem Objekterkennungsverfahren vorgeschlagen. Ein Beispiel hierfür findet sich bei Miau und Itti [103]. Bei diesem wird der Bottom-Up-Salienzdetektor nach Itti et al. [57] eingesetzt. Dieser detektiert saliente Punkte, die anschließend von einem Objekterkennungsverfahren ausgewertet werden. Das System wurde mit synthetischen Daten aus Kreisen und Rechtecken getestet und erzielte dort im Vergleich zu einer erschöpfenden Suche ähnlich

gute Ergebnisse, jedoch mit einem erheblichen Geschwindigkeitsvorteil. Das Verfahren wurde von Walther et al. [89] um eine Region-Grow-Methode erweitert, die den jeweils betrachteten salienten Punkt als Startpunkt verwendet (vgl. Abschnitt 5.1.3). Es wird dabei jeweils auf der Merkmalskarte segmentiert, die den stärksten Beitrag zum salienten Punkt liefert. Der segmentierte Bereich wird anschließend von einem Klassifizierer ausgewertet. Aufbauend auf diesem Konzept vergleichen Rutishauser et al. [90] ein Objekterkennungsverfahren auf Basis eines Suchrasters mit einem Verfahren, das einen Salienzdetektor einsetzt. Die Klassifizierung erfolgt in beiden Fällen durch die Auswertung von SIFT-Merkmalen [14]. Für ihre Experimente erstellen die Autoren Bildmontagen aus jeweils einem Objekt und einem Hintergrundbild in unterschiedlichen Größenverhältnissen. Die Ergebnisse hierfür zeigen, dass bei den größten Objekten die Rate der korrekten Treffer bei Einsatz der Salienzdetektion um ca. 10% und bei den kleinsten Objekten um ca. 50% besser ist, als beim Suchrasterverfahren. Bei einer weiteren Untersuchung ersetzen die Autoren den Klassifizierer durch einen menschlichen Probanden. Ist dieser in der Lage, das vom Salienzdetektor aus dem Kontext herausgelöste Segment dem gesuchten Objekt zuzuordnen, wird dies als Treffer des Salienzdetektors gewertet. Aus den hieraus erzielten Ergebnissen schließen die Autoren, dass der Einsatz aufmerksamkeitsbasierter Methoden bei Objekterkennungsproblemen grundsätzlich Vorteile bringen kann.

Kokkinos et al. [104] schlagen ein zweistufiges Verfahren vor, bei dem im ersten Schritt in einem Bottom-Up-Prozess saliente Punkte im Bild ermittelt werden. Aus Kacheln um diese Punkte werden Objekthypothesen aufgestellt. Im anschließenden Top-Down-Prozess werden im Rahmen eines teilebasierten Objekterkennungsverfahrens (vgl. Abschnitt 3.3) die Kacheln nacheinander ausgewertet. Dabei wird in der Umgebung gezielt nach noch fehlenden Objektteilen gesucht, die in der jeweiligen Objekthypothese fehlen.

Avraham und Lindenbaum [105] stellen ebenfalls einen Ansatz vor, bei dem in einem Bottom-Up-Verfahren eine Reihe von Objekthypothesen aufgestellt und anschließend mittels Top-Down-Erkennung ausgewertet werden. Die Besonderheit des Ansatzes ist, dass sich die Reihenfolge, in der die Objekthypothesen bearbeitet werden, abhängig von den vorangegangen Ergebnissen noch nachträglich ändern kann. Die Annahme dabei ist, dass innerhalb einer Szene ähnliche Regionen einen gleichen oder den selben Ursprung haben. Ergibt also die Top-Down-Auswertung einer Region, dass diese relevant oder nicht relevant ist, kann diese Information genutzt werden, um die Priorität ähnlicher Regionen festzulegen.

Tagare et al. [106] schlagen ebenfalls ein zweistufiges Verfahren vor. Bei dem Aufmerksamkeitsmodell handelt es sich hier jedoch um einen aufgabenorientierten Top-Down-Ansatz. Im ersten Schritt werden auf der gesamten Bildfläche Merkmalsdeskriptoren extrahiert, welche sich aus Farb- und Orientierungsmerkmalen zusammensetzen. Anschließend folgt ein iterativer Prozess. Bei jeder Iteration wird der Merkmalsdeskriptor ausgewählt, für den die größte Wahrscheinlichkeit

besteht, dass er zum Zielobjekt gehört. Hierfür wird ein Maximum-Likelihood-Schätzer (Bishop [12], S. 140 ff.) eingesetzt. Im zweiten Schritt wird der zum Deskriptor gehörende Bildbereich ausgewertet. Wird die Objekthypothese bestätigt, wird der iterative Prozess beendet. Andernfalls wird mit der nächsten Iteration fortgefahren.

Eine Strategie kann es sein, das Objekterkennungsproblem stärker in das Konzept der sakkadischen Augenbwegungen zu integrieren. Bandera et al. [107] schlagen ein Objekterkennungsverfahren vor, welches auf dem Prinzip des fovealen Sehens basiert. Die Fovea wird durch ein zur Mitte hin feiner werdendes Gitternetz modelliert. Die Position des Gitternetzes in der Szene wird durch ein Aufmerksamkeitsmodell von Punkt zu Punkt verschoben. An jeder Position werden Merkmale im fovealen Bereich extrahiert und anschließend klassifiziert. Bei der Klassifizierung werden vorangegangene Ergebnisse einbezogen. Das bedeutet, es wird überprüft, ob bereits durch andere Fixationspunkte auf ein bestimmtes Objekt geschlossen werden konnte, und ob dies durch die aktuelle Auswertung bestätigt werden kann.

Paletta et al. [108] [109] schlagen ein Verfahren vor, bei dem die sakkadischen Bewegungen durch den Objekterkennungsprozess bestimmt werden. Dies geschieht mittels verstärkendem Lernen (Bishop [12], S. 3 f.), bei dem aus dem aktuellen Zustand die nächste Aktion abgeleitet wird. Im ersten Schritt werden Keypoints detektiert und die zugehörigen SIFT-Merkmale extrahiert (siehe Abschnitt 2.3.1.3). Anschließend werden Sakkaden gebildet, indem sequentiell die Merkmale anhand des jeweiligen Zustands der Objekterkennung ausgewählt werden. Es wird stets das Merkmal ausgewählt, welches im aktuellen Zustand zu der am besten gesicherten Entscheidung führt.

5.3 VERWANDTE PROBLEMSTELLUNGEN

Ein wichtiges Forschungsgebiet für visuelle Aufmerksamkeitsmodelle ist die Roboterforschung. Ist ein Roboter mit einer Kamerasensorik ausgestattet, kommt der Einsatz entsprechender Modelle in Betracht. In Kapitel 3.1 wurde das Thema der Objekterkennung behandelt. Alle dort betrachteten Problemstellungen bezogen sich auf Einzelbildaufnahmen. Die Problembetrachtung erweitert sich bei Robotern dahingehend, dass sie über eine Aktorik zur Navigation und Interaktion in ihrer Umgebung verfügen.

Ein Beispiel für eine Problemstellung aus dem Bereich der Roboterforschung ist die Suche nach einem bekannten Objekt in einer komplexen Umgebung. Minut und Mahadevan [8] schlagen einen aufmerksamkeitsbasierten Ansatz für die visuelle Objektsuche vor. Sie verwenden eine Kamera mit Zoomfunktion, die über zwei Achsen gedreht werden kann. Mit dieser werden sakkadische Bewegungen analog zu menschlichen Wahrnehmung modelliert. Foveales Sehen wird durch starkes Hineinzoomen simuliert. Die sakkadischen Bewegungen werden

fortgeführt, bis das Objekt an einem Fixationspunkt gefunden wurde. Letzteres wird durch einen Abgleich mit einer Bildvorlage des gesuchten Objekts bestimmt. Die Sakkaden werden auf Basis des Merkmalsintegrationsmodells von Itti et al. [57] bestimmt. Diese werden durch verstärkendes Lernen (Bishop [12], S. 3 f.) optimiert, indem frühere Aufenthaltsorte des Objekts im Entscheidungsprozess einbezogen werden.

Ein weiteres Beispiel zur Objektsuche findet sich bei Shubina und Tsotsos [9]. Sie verwenden einen mobilen Roboter auf vier Rädern (Pioneer 3). Auf diesem ist eine Stereokamera ohne Zoomfunktion montiert, die über zwei Achsen bewegt werden kann. Sie setzen das Aufmerksamkeitsmodell von Frintrop et al. [110] [111] ein. Dieses erweitert das Bottom-Up-Modell von Itti et al. [57] um zusätzliche Top-Down-Elemente. Das verwendete Modellwissen umfasst Merkmale des Zielobjekts. Es wird zur Gewichtung der Merkmalskarten genutzt (vgl. Abbildung 5.1.8 in Abschnitt 5.1.4). Dabei werden Merkmale, die eher zum Zielobjekt gehören, angehoben, während Bereiche, die eher zum Hintergrund gehören, abgeschwächt werden. Das Suchverfahren arbeitet in zwei Phasen. In der ersten Phase werden Fixationspunkte bestimmt, aus denen Regionen extrahiert werden. Dies werden mit dem gesuchten Objekt abgeglichen. Im nächsten Schritt werden Bewegungsaktionen ausgeführt. Diese sollen entweder eine bessere Sicht auf vielversprechende Inhalte ermöglichen, oder die Suche auf andere Bereiche der Umgebung lenken.

Eine andere Problemstellung aus dem Bereich der Roboterforschung ist die Erkundung der Umgebung nach neuen, unbekannten Objekten (engl.: object discovery). Hierbei geht um das unüberwachte Lernen von Objektrepräsentationen, die zu einem späteren Zeitpunkt genutzt werden können, um das gelernte Objekt wiederzuerkennen. Ein Verfahren hierzu findet sich bei Forssén et al. [81] [112]. Die Autoren verwenden einen mobilen Roboter, der zur Modellierung des peripheren und fovealen Sehens über eine Kamera mit niedriger und einer weiteren Kamera mit hoher Auflösung verfügt. Es wird zunächst der in Abschnitt 5.1.2 beschriebene Salienzdetektor von Hou und Zhang [84] verwendet, um saliente Objekte mit der peripheren Kamera zu detektieren. Dies geschieht in einem vorgegebenen Szenario, in dem der Roboter die Objekte mittels Abstandssensoren umfahren kann. Beim Umfahren wird die foveale Kamera eingesetzt, um hochauflösende Bilder des salienten Objekts aufzunehmen. Aus diesen wird schließlich eine BoF-Repräsentation (siehe Abschnitt 3.4) des Objekts erstellt.

Eine Variante der Objekterkundung abseits der Robotik ist die Verwendung einer Bilddatenbank, in der es gilt, unbekannte Objekte zu detektieren und ähnliche Objekte zu gruppieren. Ein entsprechender Ansatz hierzu findet sich bei Zhu et al. [113] [114]. Die Autoren verwenden einen Salienzdetektor auf Basis eines Suchfensters (vgl. 5.1.6 in Abschnitt 5.1.3). Aus den salientesten Regionen in jedem Bild werden Objektrepräsentation extrahiert. Diese werden anschließend im Rahmen eines Clusterverfahrens in Gruppen eingeteilt.

6 KONZEPTION UND UMSETZUNG

Einleitend in Kapitel 1 wurden bereits die Zielsetzungen für den praktischen Teil dieser Arbeit erläutert. Auf Basis von Methoden der objektbasierten Aufmerksamkeit soll ein System zur Lokalisierung interessanter Objekte erstellt werden. Gehört das interessante Objekte einer bekannten Objektkategorie an, soll dieses auch entsprechend gelabelt werden. Andernfalls wird es als unbekanntes Objekt eingestuft. Das System ist als Alternative zu einer erschöpfenden Suche über den gesamten Bildbereich zu verstehen, bei dem im Verhältnis zu den tatsächlich vorhandenen Objekten im Bild eine viel größere Anzahl an Bildausschnitten ausgewertet werden müsste. Dies kann potentiell zu einer hohen Anzahl an Fehldetektionen führen, wenn bestimmte Objekte in eigentlich unbedeutenden Hintergrundflächen erkannt werden. Diese Gefahr erhöht sich tendenziell bei einer zunehmenden Anzahl an betrachteten Objektkategorien, da ein größerer Lösungsraum mehr Möglichkeiten für Fehldetektionen bietet. Der Einsatz eines Salienzdetektors soll ermöglichen, sich auf wenige saliente Bereiche zu konzentrieren, um so eine große Anzahl irrelevanter Bildbereiche mit vergleichsweise geringem Rechenaufwand von vornherein auszuschließen. Auf diese Weise arbeitet das Detektionssystem erheblich zeiteffizienter, da die Bestimmung der Proto-Objekte deutlich weniger Rechenleistung in Anspruch nimmt, als die Klassifizierung der eingesparten Bereiche es tun würde.

Aufbauend auf den Ausführungen der vorangegangenen Kapiteln wird nun in diesem Kapitel die Konzeption und praktische Umsetzung für das aufmerksamkeitsbasierte Objekterkennungssystem beschrieben. In Abschnitt 6.1 werden zunächst die wichtigen Vorüberlegungen erörtert, die den präsentierten Methoden zugrunde liegen. In Abschnitt 6.2 wird die Übersicht zum Gesamtsystem betrachtet. In den darauffolgenden Abschnitten werden die einzelnen Teilaspekte im Detail betrachtet. Abschnitt 6.3 beschreibt das Segmentierungsverfahren, das im Rahmen der Salienzdetektion eingesetzt wird. Abschnitt 6.4 beschreibt den Salienzdetektor selbst. Schließlich wird in Abschnitt 6.5 das Vorgehen bei der Klassifizierung erläutert.

6.1 VORÜBERLEGUNGEN

In Kapitel 1 wurde entsprechend Abbildung 1.2.1 für die Umsetzung der aufmerksamkeitsbasierten Objekterkennung ein zweistufiges Modell vorgesehen, welches zunächst saliente Bereiche detektiert und diese anschließend klassifiziert. Dieses Vorgehen soll im Folgenden nun näher konkretisiert werden. Analog zur

menschlichen Aufmerksamkeit, die sich in eine präattentive und eine attentive Phase unterteilt, soll ein zweistufiger Prozess umgesetzt werden, der sich aus einem generischen und einem problemspezifischen Teil zusammensetzt. Für den generischen Teil sollen ausschließlich Bottom-Up-Methoden eingesetzt werden, um flexible Anwendungsmöglichkeiten zu gewährleisten. Dies bedeutet insbesondere, dass keine Beispielbilder verwendet werden sollen, um bspw. die Verteilung bestimmter Parameter zu bestimmen. Beispielbilder können immer nur einen Ausschnitt der Realität wiedergeben und verzerren diese zwangsläufig. Ihr Einsatz eignet sich daher eher für eingegrenzte Problemstellungen, die durch den jeweiligen Datensatz gut abgebildet werden. Aus den gleichen Gründen sollen auch sonst keine Annahmen getroffen werden, die sich nicht auf natürliche Bilder im Allgemeinen anwenden lassen. Da das zu entwerfende Verfahren auch auf schwierigen Eingabebildern möglichst gute Ergebnisse erzielen soll, stellt diese Einschränkung das zentrale Problem dar, das bei der Umsetzung adressiert wird. Ergänzend wird in Abschnitt 6.4.4 die Möglichkeit betrachtet, im Sinne einer gerichteten Aufmerksamkeit Vorwissen in den Prozess einfließen zu lassen.

Der Anspruch, ein reines Bottom-Up-Verfahren zu realisieren, schließt Methoden aus, wie sie in Abschnitt 5.1.4 beschrieben wurden, da sie auf den Einsatz von Modellwissen beruhen. Es wird stattdessen lediglich von zwei generellen Eigenschaften ausgegangen, die hier für Objekte im Allgemeinen angenommen werden. Zum einen wird angenommen, dass Objekte salient sind. Zum anderen wird per Definition davon ausgegangen, dass Objekte eine geschlossene Einheit darstellen und entsprechend auch eine äußere Kontur aufweisen (vgl. Abschnitt 3.2). Der Ansatz der Redundanzreduktion, der in Abschnitt 5.1.2 betrachtet wurde, wird hier nicht verfolgt, da dieser lediglich auf die Herausstellung globaler Alleinstellungsmerkmale im Bild abzielt. Dieses Vorgehen funktioniert gut bei vergleichsweise einfachen Szenen, bei denen sich ein Objekt vor einem weitestgehend zusammenhängenden Hintergrund abhebt. Bei schwierigeren Konstellationen wird dieser Ansatz jedoch der Komplexität des Problems nicht gerecht. Hier sollte auch die Stellung eines Objektkandidaten zu seiner lokalen Umgebung berücksichtigt werden. Diese Überlegung führt zu dem Ansatz, zunächst mögliche Objektkandidaten auszumachen und anschließend deren Salienz im Verhältnis zum lokalen Umfeld zu betrachten. Dies schließt den Ansatz aus, zunächst eine Salienzkarte zu ermitteln und erst im Nachgang auf Basis von dieser nach Objektkonturen zu suchen, wie es in Abbildung 5.1.5 in Abschnitt 5.1.3 illustriert wurde. Die Suchfenstermethode nach Abbildung 5.1.6 käme nach den bisherigen Überlegungen zwar in Frage, ist für die verfolgte Bottom-Up-Strategie aber eher problematisch. Bei dieser Methode werden Objektkandidaten an sehr vielen Bildpositionen extrahiert. Ohne statistische Annahmen bspw. über Kontrast, Helligkeit und Kantenstärken, oder auch typische Positionen, Größen und Formen von Objekten, fällt es bei einem großen Pool von Objektkandidaten tendenziell schwierig, diese nach ihrer Güte zu ordnen. Ein besserer Ansatz ist es

daher, sich mit einem Blick auf die Gesamtstruktur des Bildes auf vergleichsweise wenige aber dafür vielversprechende Objektkandidaten zu beschränken. Diese Form der Komplexitätsreduktion können bestimmte Segmentierungsverfahren leisten. Verfahren auf Basis von Startpunkten, wie sie bei Ansätzen entsprechend Abbildung 5.1.4 eingesetzt werden, zählen jedoch nicht dazu, da sie nur jeweils einen Ausschnitt des Bildes betrachten. Eher kommen Verfahren in Betracht, die die Gesamtstruktur des Bildes berücksichtigen und die Bildfläche in disjunkte Segmente unterteilen. Das Verfahren von Felzenszwalb und Huttenlocher [19] leistet genau dies (siehe Abschnitt 2.2.1) und bildet hier deshalb die Ausgangslage. Der große Vorteil der Komplexitätsreduktion ist auch gleichzeitig die größte Schwäche des Segmentierungsansatzes. Objekte, die aufgrund einer Über- oder Untersegmentierung verfehlt werden, können von vornherein nicht mehr detektiert werden. Um einen möglichst guten Kompromiss aus Komplexitätsreduktion und Abdeckung der tatsächlich vorhandenen Objekte zu erzielen, wird deshalb der Ansatz eines hierarchischen Segmentierungsverfahrens verfolgt (siehe Abschnitt 2.2.2). Hierzu wird das Verfahren nach Felzenszwalb und Huttenlocher [19] entsprechend erweitert. Das genaue Vorgehen hierbei wird in Abschnitt 6.3 erläutert.

Nachdem im ersten Schritt Objektkandidaten segmentiert wurden, sollen nun im zweite Schritt deren Salienz bewertet werden. Dies entspricht der Strategie, die in Abschnitt 5.1.3 in der Abbildung 5.1.7 illustriert wurde. In diese Betrachtung sollen Aspekte der menschlichen Aufmerksamkeit einfließen, die durch Verhaltensexperimente gut belegt sind. Als solche werden hier das Prinzip der Merkmalsintegration, die Zentrum-Umgebung-Hypothese sowie die Winner-Take-All-Strategie identifiziert (siehe Abschnitt 4.1). Des Weiteren sollen die gleichen Salienzmerkmale bezüglich Farbe, Helligkeit und Struktur betrachtet werden wie bei Itti et al. [57], da diese sich stark an der menschlichen Wahrnehmung orientieren (siehe Abschnitt 4.3). Anstatt nur einzelne saliente Merkmale der Objektkandidaten zu betrachten (bspw. [103] [89] [3]), wird die Salienz von Regionen als Ganzes ermittelt (bspw. [86] [96]). Dieser Ansatz passt gut in die objektbasierte Strategie, da davon ausgegangen wird, dass sich die interessanten Objekte als Ganzes von ihrer Umgebung abheben. Im ersten Schritt werden hierfür lokale Salienzmerkmale extrahiert, um aus diesen im zweiten Schritt regionale Merkmale zu ermitteln. Hierfür muss eine entsprechende Vorschrift definiert werden. Eine weitere Vorschrift wird benötigt, um die visuelle Distanz zweier Regionen bezüglich ihrer regionalen Merkmale zu messen. In Abschnitt 6.4.2 werden hierzu verschiedene Varianten unterschiedlicher Komplexität erörtert. Die Distanzwerte der unterschiedlichen Merkmalstypen sollen schließlich zu einem einzigen Salienzwert integriert werden. Anders als bei verschiedenen Top-Down-Methoden, bei denen die Integration verschiedener Merkmale mittels Modellwissen optimiert wird (bspw. [95]), soll hier auf Wertenormalisierungen und Linearkombinationen nach einer einheitlichen Vorschrift gesetzt werden

(vgl. Abschnitt 5.1.1), da hierfür kein Vorwissen erforderlich ist. Nachdem die Salienzwerte aller Regionen ermittelt wurden, werden entsprechend der Winner-Take-All-Strategie nacheinander die salientesten Regionen als Proto-Objekte ausgewählt und klassifiziert.

Bei der Klassifizierung der Proto-Objekte wird auf bewährte Methoden gesetzt. Dies sind im Einzelnen der Bag-Of-Features-Ansatz (siehe Abschnitt 3.4) unter Verwendung von SIFT-Merkmalen (siehe Abschnitt 2.3.1.3) und einem Random-Forest-Klassifizierer (siehe Abschnitt 2.3.2.1). Der Bag-Of-Features-Ansatz hat den Vorteil, dass er sich anders als bspw. teilebasierte Ansätze (siehe Abschnitt 3.3) ohne besondere Vorkehrungen auf unterschiedliche Objektkategorien anwenden lässt. Random-Forest-Klassifizierer haben den Vorteil, dass sie gut mit großen, schwachbesetzten Merkmalsvektoren skalieren, wie sie beim Bag-Of-Features-Ansatz typischerweise auftreten.

6.2 Übersicht des Gesamtsystems

Abbildung 6.2.1 zeigt eine Übersicht des Gesamtsystems. Entsprechend der Ausführungen im letzten Abschnitt, basiert der Salienzdetektor, der auf der linken Seite der Abbildung dargestellt ist, auf dem objektbasierten Ansatz nach Abbildung 5.1.7. Die Bildfläche wird zunächst mittels eines Segmentierungsverfahrens in Regionen unterteilt. Das Segmentierungsverfahren verfolgt einen multiskalaren, hierarchischen Ansatz und wird in Abschnitt 6.3 betrachtet. Parallel zur Segmentierung werden die lokalen Salienzmerkmale extrahiert, die auf Helligkeit, Farbe und Orientierung basieren. Die Extraktion erfolgt ebenfalls auf mehreren Skalierungsstufen. Im nächsten Schritt werden auf Basis der lokalen Merkmale regionale Merkmale ermittelt. Unter einem regionalen Merkmal ist ein Merkmal zu verstehen, das sich auf eine gesamte Region bezieht. Eine genaue Beschreibung der Merkmalsextraktion erfolgt in Abschnitt 6.4.1. Als nächstes erfolgt die Bestimmung der regionalen Salienzen. Dabei wird jeder Region ein Salienzwert zugeordnet. Das Vorgehen dabei beruht auf den Prinzipien der Merkmalsintegration und der Zentrum-Umgebungs-Hypothese. Eine Beschreibung hierzu erfolgt in Abschnitt 6.4.2. Anhand der regionalen Salienzwerte wird

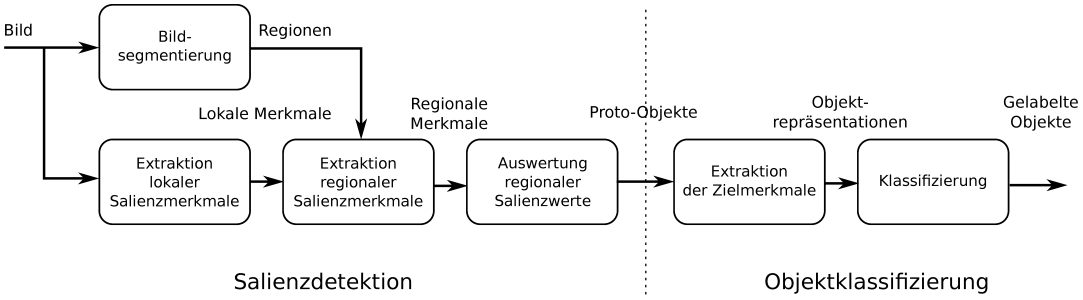

Abbildung 6.2.1: Übersicht des Objekterkennungssystems.

schließlich eine Liste von Proto-Objekten erstellt. Diese sind für die Objektklassifizierung bestimmt, die auf der rechten Seite der Abbildung dargestellt ist. Die Merkmalsextraktion erfolgt auf Basis des Bag-Of-Features-Ansatzes und wird in Abschnitt 6.5.1 beschrieben. Klassifizierung und Training basieren auf dem Random-Forest-Verfahren und werden in Abschnitt 6.5.2 betrachtet.

6.3 SEGMENTIERUNG

Das Segmentierungsverfahren soll die gesamte Bildfläche in Regionen aufteilen, aus denen dann später die salientesten ausgewählt werden. Entsprechend dem objektbasierten Ansatz wird die Segmentierung als integrativer Teil der Salienzdetektion aufgefasst. Hieraus ergibt sich die Anforderung, im Rahmen der Salienzdetektion zwecks einer effizienten Verarbeitung nur grundlegende Merkmale einzusetzen. Es werden entsprechend einfache Kantenmerkmale betrachtet. Auf die Verwendung komplexerer Merkmale oder den Einsatz von Modellwissen wird hingegen verzichtet.

Ein weiterer wesentlicher Punkt ist die Verwendung eines multiskalaren, hierarchischen Segmentierungsmodells. Durch ein solches Modell wird vermieden, sich von vornherein auf eine feste Bildaufteilung festzulegen. Dieser Ansatz soll ermöglichen, Objekte mit stark unterschiedlichen Größen zu detektieren. Bei Verwendung eines einstufigen Modells führt eine feine Parametrierung dazu, dass bei großen Objekten eher Teilbereiche von diesen segmentiert werden, während bei einer gröberen Ausrichtung kleinere Objekte übergangen werden. Bei dem hier verfolgten multiskalaren Ansatz wird zunächst mit einer feinen Segmentierung begonnen. Diese bildet die Ausgangslage für die nächstgröbere Skalierungsstufe, bei der kleine Teilbereiche zu größeren Bereichen zusammengefügt werden.

Ein weiteres Problem, das hier adressiert wird, ist eine wechselnde Qualität der Segmentierungsergebnisse von Bild zu Bild aufgrund von variierenden Faktoren wie Kontrast, Fokus und Lichtverhältnisse. Um diesem Problem zu begegnen, wird eine adaptive Anpassung an das Eingabebild vorgenommen. Die Zielgröße, an der sich die Adaption ausrichtet, ist die Anzahl der Segmente. Für jede Skalierungsstufe wird eine maximale und eine minimale Anzahl an Regionen festgelegt, die nicht über- bzw. unterschritten werden darf. Auf diese Weise lässt sich sicherstellen, dass die Größe des Kandidatenpools, aus dem die salienten Regionen ausgewählt werden, innerhalb eines sinvollen Rahmens liegt.

Abbildung 6.3.1 zeigt eine Übersicht des Segmentierungsverfahrens. Von links nach rechts wird die jeweils nächste Skalierungsstufe verarbeitet. Die Anzahl der Stufen kann variabel gestaltet werden und lässt sich fortführen, solange noch mehr als ein Segment vorhanden ist. Nach einer initialen Glättung des Eingabebildes zur Reduktion hochfrequenter Rauscheinflüsse, wird jeweils zwischen zwei Skalierungsstufen entsprechend dem in Abschnitt 2.1.1 beschriebenen Scale-Space-Prinzip eine weitere Glättung und eine Unterabtastung durchge-

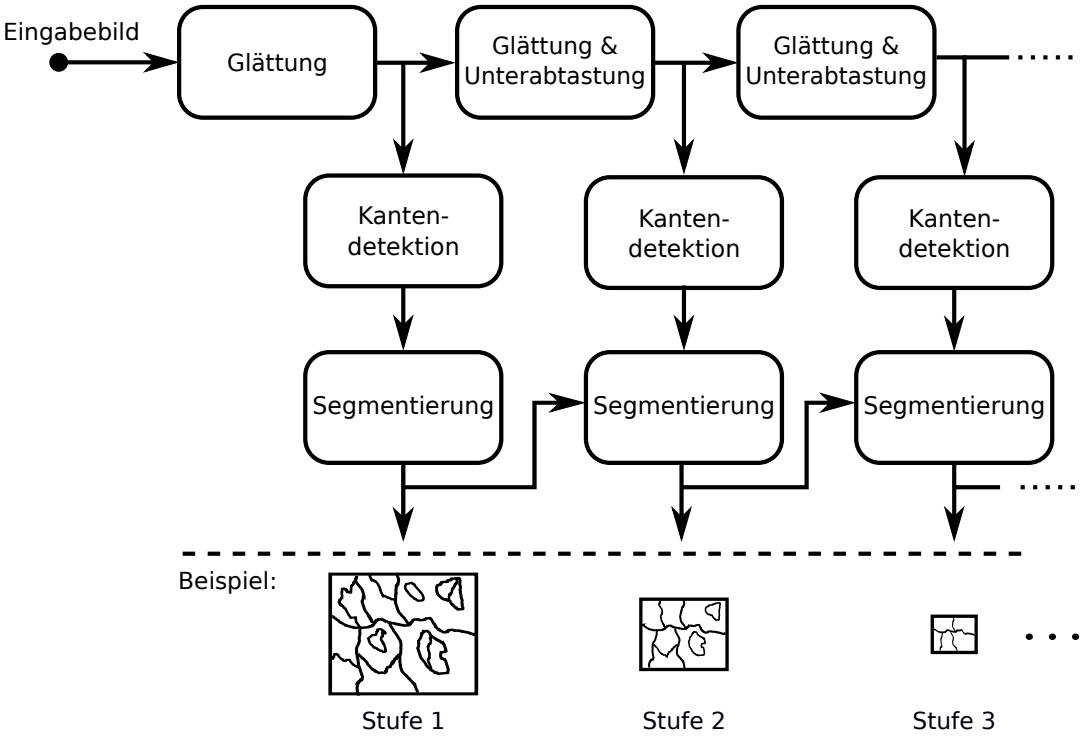

Abbildung 6.3.1: Schema der multiskalaren Segmentierung.

führt, bei der die Bilddimensionen halbiert werden. Zur Glättung wird jeweils ein Gauß-Filter eingesetzt. Die hochfrequenten Kanten werden auf diese Weise von Stufe zu Stufe reduziert und bei der Segmentierung entsprechend nicht mehr berücksichtigt.

Die eigentliche Segmentierung basiert auf dem in Abschnitt 2.2.1 betrachteten Graphenverfahren von Felzenszwalb und Huttenlocher [19], wird jedoch zwecks einer multiskalaren Betrachtung wie folgt erweitert. Auf jeder Skalierungsstufe, l, werden die Kantenmerkmale in horizontaler, vertikaler und diagonaler Richtung, $d(I_l(j,i), I_l(j,i+1))$, $d(I_l(j,i), I_l(j+i,i))$ und $d(I_l(j,i), I_l(j+1,i+1))$ bestimmt, wobei sich die Distanz zweier benachbarter Bildpunkte, $p = (p_r, p_g, p_b)$ und $q = (q_r, q_g, q_b)$, nach wie vor nach euklidischer Norm bestimmt:

$$d(p,q) = \sqrt{(p_r - q_r)^2 + (p_g - q_g)^2 + (p_b - q_b)^2}. \tag{6.3.1}$$

Bei der Initialisierung der ersten Skalierungsstufe wird jeder Pixel als ein Knoten aufgefasst. Nach der Segmentierung auf einer Stufe wird jeder minimale Spannbaum zu einem Knoten zusammengefasst, der dann ein Segment repräsentiert. Der so entstehende Graph stellt gleichzeitig die Initialisierung für die nächste Stufe dar. Dabei ist zu beachten, dass die Ränder der Segmente durch die Unterabtastung geglättet werden. Das Vorgehen wird in Abbildung 6.3.2 anhand eines Beispielschemas illustriert. Die innere Differenz einer Region bestimmt sich nach wie vor anhand des maximalen Gewichts des minimalen Spannbaums. Bei der

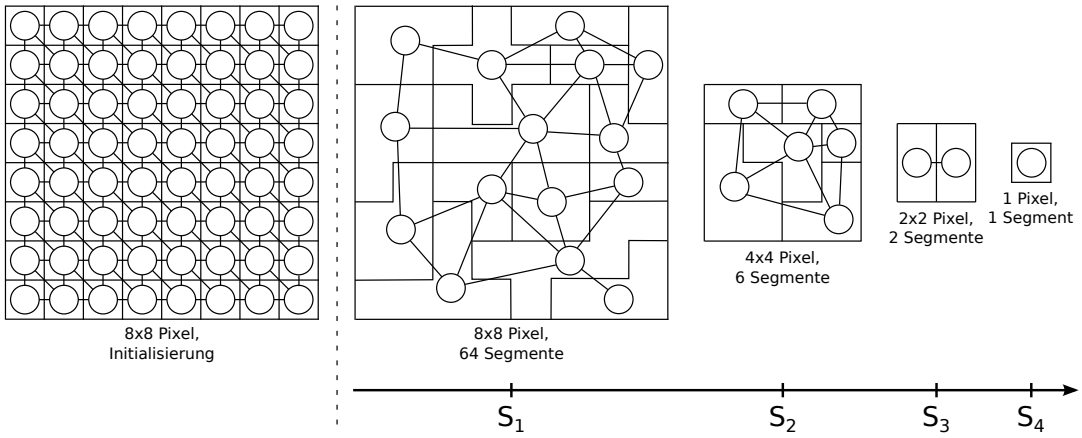

Abbildung 6.3.2: Beispielschema der multiskalaren Segmentierung.

Ermittlung der Kantengewichte muss jedoch berücksichtigt werden, dass zwei Segmente zu Beginn nun durch mehrere angrenzende Bildpunkte verbunden sein können. Deshalb werden die Kantengewichte für benachbarte Segmente durch eine Mittelung der lokalen Kantendifferenzen aller angrenzender Bildpunkte ermittelt. Für zwei angrenzende Segmente, S_n und S_m, sei B_{S_n,S_m} die Menge aller Koordinatenpaare (p,q) mit p und q angrenzend, sowie $p \in S_n$ und $q \in S_m$. Das Kantengewicht ergibt sich dann zu

$$G_{S_n,S_m} = \frac{1}{|B_{S_n,S_m}|} \sum_{\forall (p,q) \in B_{S_n,S_m}} d(I_l(p), I_l(q)). \tag{6.3.2}$$

Das weitere Vorgehen erfolgt analog zum Verfahren von Felzenszwalb und Huttenlocher [19]. Das bedeutet, die Kanten werden zunächst entsprechend ihrem Gewicht sortiert und in nicht absteigender Reihenfolge durchlaufen. Für jede Kante wird entschieden, ob diese eine segmenttrennende Kante oder eine texturelle Kante darstellt. Im letzteren Fall werden die beiden beteiligten Segmente vereint. Wie in Abschnitt 2.2.1 beschrieben wurde, entstehen so minimale Spannbäume, die das jeweilige Segment aufspannen. Auch die Schwellwerte bestimmen sich nach wie vor zu

$$T_S = G_{S,max} + \frac{c}{|S|}, \tag{6.3.3}$$

jedoch ist c hier keine Konstante, sondern wird dynamisch angepasst, um die Anzahl der Segmente zu regulieren. Hierzu wird der Durchlauf über alle Kanten solange wiederholt, bis die resultierende Segmentzahl innerhalb des gewünschten Bereichs liegt. Nach jedem Durchlauf wird c mit einem Faktor, $a > 1$, angepasst:

$$c_T = \begin{cases} a c_{T-1}, \text{ bei zu wenigen Segmenten} \\ \frac{1}{a} c_{T-1}, \text{ bei zu vielen Segmenten.} \end{cases} \tag{6.3.4}$$

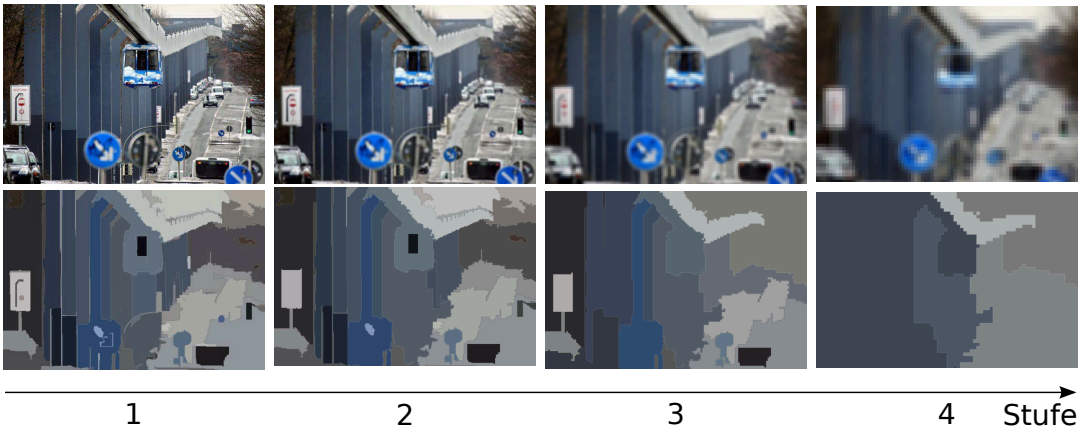

Abbildung 6.3.3: Beispiel der multiskalaren Segmentierung mit vier Stufen. Oben: Scale-Space. Unten: Die zugehörigen Segmentbilder.

Da die Sortierung der Kanten auch bei mehreren Durchläufen nur einmal durchgeführt werden muss, steigt die Laufzeit des Verfahrens bei nicht zu kleine Werten für a nicht sehr stark an. Des Weiteren kann bei Videosequenzen der letztendlich ermittelte Wert für c als Prädiktion für das nächste Eingabebild verwendet werden.

Im Ergebnis ist auf jeder Skalierungsstufe, l, die Bildfläche in eine Menge disjunkter Segmente, $\mathcal{S}_l$, eingeteilt. Als Nachbearbeitungsschritt werden Segmente, die zu klein sind, um aussagekräftige Informationen enthalten zu können, verworfen. Hierzu wird ein geeigneter Größenschwellwert festgelegt. Abbildung 6.3.3 zeigt ein Beispiel mit vier Skalierungsstufen. Die obere Reihe zeigt den Scale-Space des Beispielbildes mit vier Skalierungsstufen. Unter den einzelnen Bildern ist jeweils das zugehörige Segmentbild dargestellt. Zur Darstellung wird der über der jeweiligen Region gemittelte Farbwert verwendet. Die Ränder der Segmente sind etwas hervorgehoben, um die Übergänge zwischen den Segmenten zu betonen. Das Beispiels zeigt, wie die Zahl der Segmente von Stufe zu Stufe abnimmt.

Für das Verfahren müssen folgende Parameter konfiguriert werden. Zunächst ist die Varianz des Gauß-Filters zu wählen. Hier kann der gleiche Wert gewählt werden, wie bei der Extraktion der lokalen Merkmale (vgl. Abschnitt 6.4.1). Ein weiterer Parameter ist der Faktor a, der bei der Berechnung des dynamischen Schwellwerts einfließt und einen Einfluss auf die Laufzeit des Verfahrens hat. Des Weiteren muss für die jeweilige Skalierungsstufe eine maximale und eine minimale Anzahl an Segmenten festgelegt werden. Der letzte Parameter ist der Größenschwellwert der Nachbearbeitung. Bei der Wahl dieses Wertes geht es jedoch lediglich darum, sehr kleine Segmente mit wenigen Pixeln zu unterdrücken.

6.4 SALIENZDETEKTION

Der im Folgenden vorgestellte Salienzdetektor basiert auf dem Prinzip der objektbasierten Aufmerksamkeit. Es wird bewertet, wie stark sich abgegrenzte Bildregionen von ihrem Umfeld unterscheiden. Die entsprechende Eigenschaft wird hier als *visuelle Distanz* bezeichnet. Beim objektbasierten Ansatz wird von der Annahme ausgegangen, dass es sich bei salienten Regionen um Objekte handeln könnte, was durch den Begriff Proto-Objekt ausgedrückt wird (siehe Abschnitt 4.1). Der Detektor ermittelt eine Liste von Proto-Objekten. Dabei wird dem Prinzip der Aufmerksamkeitsverschiebung gefolgt. Das heißt, es wird

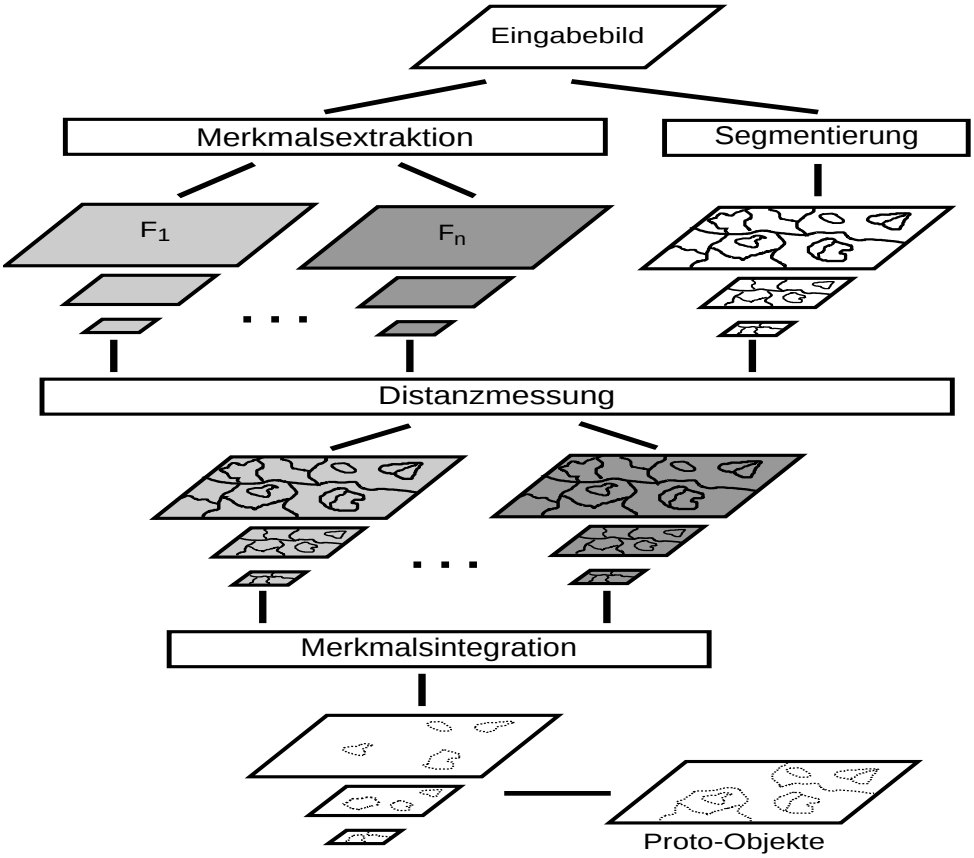

Abbildung 6.4.1: Schematische Darstellung der Salienzdetektion. Diese basiert auf dem Prinzip der Merkmalsintegration und der objektbasierten Aufmerksamkeit. Auf mehreren Skalierungsstufen wird für unterschiedliche Merkmalstypen die visuelle Distanz bestimmt. Es erfolgt die Merkmalsintegration, bei der die Distanzwerte zu einem einzigen Salienzwert pro Region zusammengefasst werden. Hieraus kann dann schließlich eine Liste mit Proto-Objekten erstellt werden.

davon ausgegangen, dass die Proto-Objekte sukzessiv in nicht-aufsteigender Reihenfolge nach dem Grad ihrer Salienz angesteuert werden.

Abbildung 6.4.1 zeigt eine schematische Darstellung des Salienzdetektors. Die einzelnen Teile werden in den weiteren Unterabschnitten im Detail erläutert. Links-oben im Bild ist die Merkmalsextraktion dargestellt. Hierbei wird dem Ansatz der Merkmalsintegration gefolgt (siehe Abschnitt 4.3). Das bedeutet, es werden unterschiedliche Merkmalstypen betrachtet, die im Prozess der Merkmalsintegration um Aufmerksamkeit konkurrieren. Die unterschiedlichen Merkmalstypen werden für mehrere Skalierungsstufen extrahiert, sodass für jeden Typ eine Merkmalspyramide entsteht. Details hierzu werden in Abschnitt 6.4.1 erläutert. Rechts-oben ist das Segmentierungsverfahren dargestellt. Wie im letzten Abschnitt bereits beschrieben wurde, erfolgt die Segmentierung ebenfalls auf mehreren Skalierungsstufen. Aus den Regionen und Merkmalen werden die visuellen Distanzen bestimmt. Pro Region und Merkmalstyp wird jeweils ein Distanzwert ermittelt. Das Vorgehen hierzu wird in Abschnitt 6.4.2 beschrieben. Die Distanzwerte werden anschließend integriert, sodass jeder Region ein Salienzwert zugeordnet wird. Hieraus kann dann schließlich eine Liste salienter Proto-Objekte erstellt werden. Dieses Vorgehen ist in Abschnitt 6.4.3 beschrieben.

6.4.1 *Merkmalsextraktion*

Abbildung 6.4.2 illustriert das Vorgehen bei der Merkmalsextraktion. In wesentlichen Teilen wird hier dem Modell von Itti et al. [57] gefolgt. Das Vorgehen dabei basiert auf dem Prinzip der Scale-Space-Repräsentation, wobei eine gröbere Skalierungsstufe durch ein Glättung und Unterabtastung erstellt wird (siehe Abschnitt 2.1). Die Merkmale werden dann auf jeder Skalierungsstufen extrahiert. Auf diese Weise entstehen eine Merkmalspyramide für die Intensität, Vier für unterschiedliche Farbmerkmale sowie Vier für unterschiedliche Orientierungsmerkmale. In der Darstellung haben die Pyramiden drei Stufen für feine, mittlere und grobe Details. Abhängig von der Größe des Eingabebildes lässt sich das Vorgehen jedoch problemlos auf eine beliebige Anzahl an Stufen erweitern. Bei drei Stufen werden insgesamt 27 Merkmalskarten erzeugt, wobei sich die Auflösung der Karten von Stufe zu Stufe halbieren. Die Intensitätsmerkmale werden

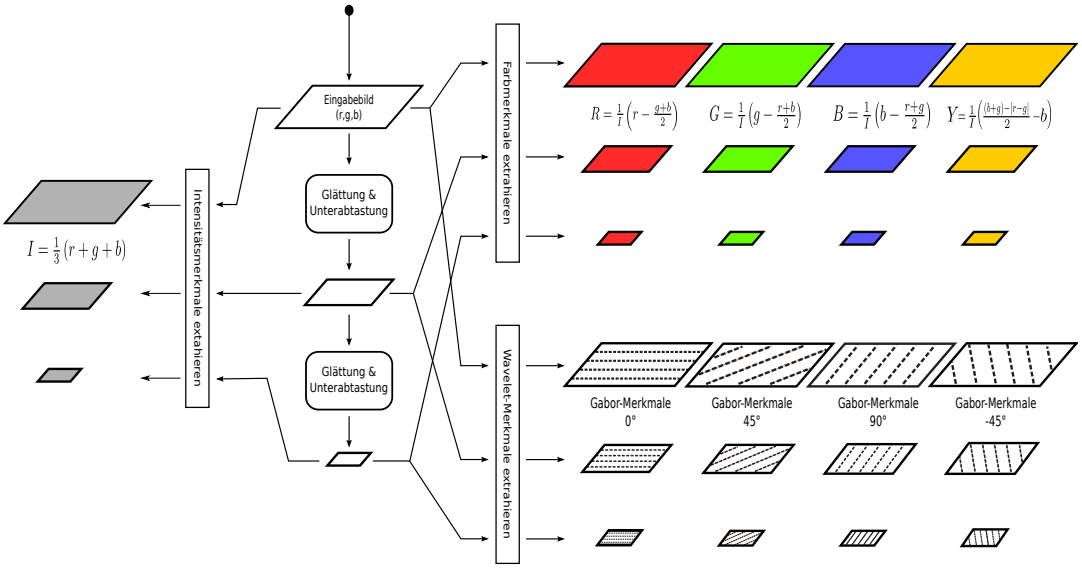

Abbildung 6.4.2: Extraktion der Salienzmerkmale. Auf mehreren Skalierungsstufen werden Merkmale extrahiert. Es entstehen so eine Pyramide für Intensität, vier für unterschiedliche Farbmerkmale und vier für unterschiedliche Orientierungsmerkmale.

entsprechend Gleichung 6.4.1 bestimmt. Des Weiteren werden die Pyramiden für Rot, Grün, Blau und Gelb entsprechend den Gleichungen 6.4.2 bis 6.4.5 erstellt.

$$I = \frac{1}{3}(r+g+b) \tag{6.4.1}$$

$$R = \frac{1}{I}\left(r - \frac{g+b}{2}\right) \tag{6.4.2}$$

$$G = \frac{1}{I}\left(g - \frac{r+b}{2}\right) \tag{6.4.3}$$

$$B = \frac{1}{I}\left(b - \frac{r+g}{2}\right) \tag{6.4.4}$$

$$Y = \frac{1}{I}\left(\frac{(b+g) - |r-g|}{2} - b\right) \tag{6.4.5}$$

Die eigentlichen Farbmerkmale ergeben sich dann anschließend aus der Rot-Grün- bzw. Blau-Gelb-Differenz:

$$C_{RG} = |R - G| \tag{6.4.6}$$

$$C_{BY} = |B - Y| \tag{6.4.7}$$

Des Weiteren werden vier Pyramiden für Strukturmerkmale generiert. Hierzu werden Gabor-Filter mit den vier Orientierungen, $0\circ$, $45°$, $90°$, $-45°$, verwendet.

Zur weiteren Adressierung der Merkmalstypen wird die Menge $\mathscr{F}$ nach folgender Formel definiert:

$$\mathscr{F} = \{I, C_{RG}, C_{BY}, 0°, 45°, 90°, -45°\}. \tag{6.4.8}$$

Nach der Extraktion der Merkmalspyramiden erfolgt eine Normalisierung der Merkmalswerte, sodass sie in einem Wertebereich zwischen Null und Eins liegen. Die Normalisierung wird nicht kartenweise sondern innerhalb einer Merkmalspyramide durchgeführt.

6.4.2 *Visuelle Distanz*

Unter der visuellen Distanz wird im Folgenden ein Größe verstanden, mit der angegeben wird, wie stark sich eine Region von anderen Regionen in ihrem Umfeld aufgrund ihrer visuellen Eigenschaften unterscheidet. Das Prinzip wurde bereits in Abschnitt 5.1.3 anhand des Modells von Cheng et al. [96] erläutert. Bei dem Ansatz werden die Eigenschaften einer Region im Verhältnis zu anderen Regionen betrachtet. Beispielsweise können strukturreiche Regionen in einem homogenen Umfeld, aber auch homogene Regionen in einem strukturreichen Umfeld hohe Distanzwerte aufweisen. Zur Messung der visuellen Distanz einer Region, r_k, wird ein Distanzmaß, D, definiert, das die Distanzen der betrachteten Region zu den anderen Regionen bestimmt und aufsummiert (vgl. Gleichung 5.1.10 in Abschnitt 5.1.3):

$$D(F, r_k) = \sum_{r_i \in \mathcal{R}_l \setminus r_k} D_r(F, r_k, r_i) w_s(r_i) w_d(r_k, r_i). \tag{6.4.9}$$

$F \in \mathscr{F}$ bezeichnet hier den betrachteten Merkmalstyp bzw. die dazugehörige Merkmalspyramide. Es wird für jeden Merkmalstyp ein eigener Distanzwert ermittelt. $\mathcal{R}_l$ bezeichnet die Menge der Regionen der l-ten Skalierungsstufe. Es werden also alle Regionen der gleichen Stufe verglichen. D_r ist eine Funktion, die bewertet, wie stark sich zwei Regionen bezüglich des betrachteten Merkmalstyps voneinander unterscheiden. Mit dem Term $w_s(r_i)$ wird die Größe der Vergleichsregion, r_i, bezeichnet. Der Unterschied zu einer großen Region wird also stärker bewertet, als zu einer kleinen Region. Dem Vorschlag von Cheng et al. [96] folgend, soll der Unterschied zu einer nah gelegenen Region stärker bewertet werden, als zu einer weiter entfernten Region. Dies wird mit dem Term $w_d(r_k, r_i)$ entsprechend folgender Formel umgesetzt:

$$w_d(r_k, r_i) = \frac{1}{d(r_k, r_i) + 1}. \tag{6.4.10}$$

Mit $d(r_k, r_i)$ wird hier der euklidische Abstand der Region r_k zur Vergleichsregion, r_i, bezeichnet. Den kürzesten Abstand zu bestimmen, wäre recht aufwändig.

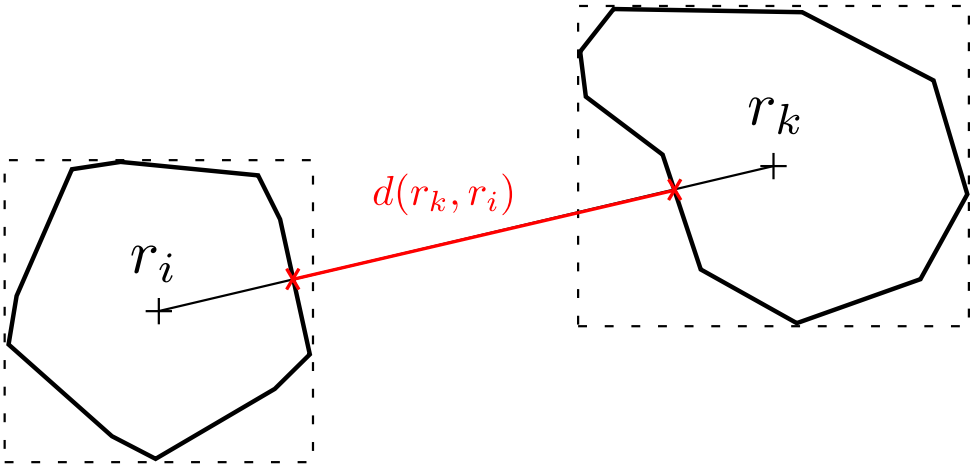

Abbildung 6.4.3: Bestimmung des Abstands zweier Regionen.

Deswegen wird hier eine Näherung verwendet, deren Prinzip in Abbildung 6.4.3 dargestellt ist. Auf der Verbindungslinie zwischen den Mittelpunkten der Bounding-Boxen werden die Grenzpunkte der Regionen mittels einer binären Suche ausfindig gemacht. Der euklidische Abstand der Grenzpunkte ergibt dann die Distanz.

Für das visuelle Distanzmaß, D_r, werden zu Vergleichszwecken drei verschiedene Varianten betrachtet, die im Folgenden beschrieben werden. Die ersten beiden Varianten basieren auf die Bildung von Histogrammen über die Merkmale der Regionen. Das Vorgehen wird in Abbildung 6.4.4 illustriert. Pro Region und Merkmalstyp wird ein Histogramm erstellt, sodass die Anzahl an Histogrammen der Anzahl an Merkmalstypen multipliziert mit der Anzahl an Regionen entspricht. Ein Histogramm wird über alle Bildpunkte einer Region ermittelt. Der Wertebereich der Merkmale wird hierzu in eine festgelegte Anzahl von N Kanälen gleicher Breite aufgeteilt. Die Histogramme werden normiert, sodass die Summe der Häufigkeiten Eins ergibt. Die visuelle Distanz zweier Regionen basiert nun auf dem Vergleich ihrer Histogramme. Viele der in der Literatur vorgeschlagenen Histogrammdistanzmaße lassen sich entweder der Klasse der Maße mit Vergleichen der Werte gleicher Kanäle (engl.: bin-by-bin distances) oder der Klasse der Maße mit kanalübergreifenden Vergleichen (engl.: cross-bin distances) zuordnen [115]. Im Allgemeinen weisen kanalübergreifende Maße im Vergleich zu Maßen mit Vergleichen gleicher Kanäle eine höhere Robustheit auf. Sie sind dafür aber auch komplexer in ihrer Berechnung.

Das erste visuelle Distanzmaß, das hier betrachtet wird, wurde von Cheng et al. [96] für die Bestimmung regionaler Salienzen vorgeschlagen. Es handelt sich

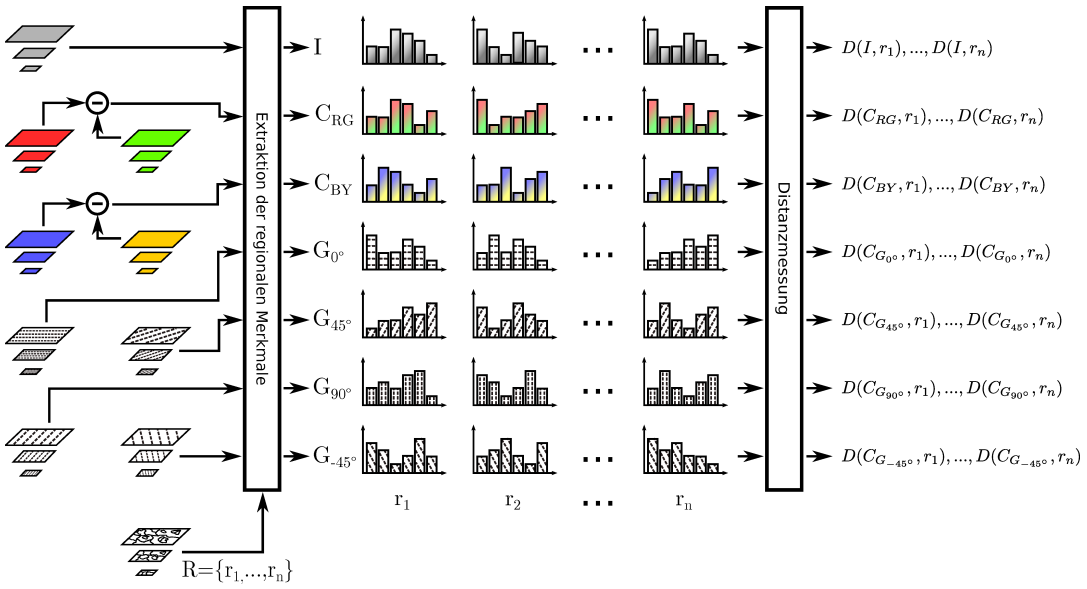

Abbildung 6.4.4: Visuelle Distanz zwischen den Regionen auf Basis von Merkmalshistogrammen. Pro Region wird für jeden der sieben dargestellten Merkmalstypen ein Histogramm extrahiert. Alle Regionen erhalten pro Merkmalstyp einen Distanzwert, der den Grad der Unterscheidung im Verhältnis zu den anderen Regionen angibt.

um ein kanalübergreifendes Histogrammdistanzmaß entsprechend der folgenden Gleichung:

$$D_r(F, r_k, r_i) = \sum_{n=1}^{N} \sum_{m=1}^{N} |(n-m) H_{F,r_k}(n) H_{F,r_i}(m)|^a. \tag{6.4.11}$$

Hierbei ist H_{F,r_k} das Histogramm für den Merkmalstyp F und der Region r_k. Es wird also der Abstand der Kanäle mit den jeweiligen Häufigkeiten multipliziert. Die Idee dabei ist es, dass Merkmale einer Region, die sich von vielen anderen Merkmalen der Vergleichsregion stark unterscheiden, einen hohen Beitrag zur Salienz leisten. Bei dieser Variante steigt die Komplexität quadratisch mit der Anzahl an Kanälen, N. Eine effiziente Berechnung ist also nur bei kleinen Werten für N möglich. Als zweite Variante wird deshalb entsprechend folgender Gleichung ein weniger komplexes Maß betrachtet, bei dem nur die korrespondierenden Kanäle verglichen werden:

$$D_r(F, r_k, r_i) = \sum_{n=1}^{N} |H_{F,r_k}(n) - H_{F,r_i}(n)|^a. \tag{6.4.12}$$

Bei diesem steigt die Komplexität lediglich linear mit der Anzahl an Kanälen an.

Die dritte der betrachteten Varianten basiert auf einer Kombination aus arithmetischem Mittel und Signalvariation, die jeweils über alle Bildpunkte

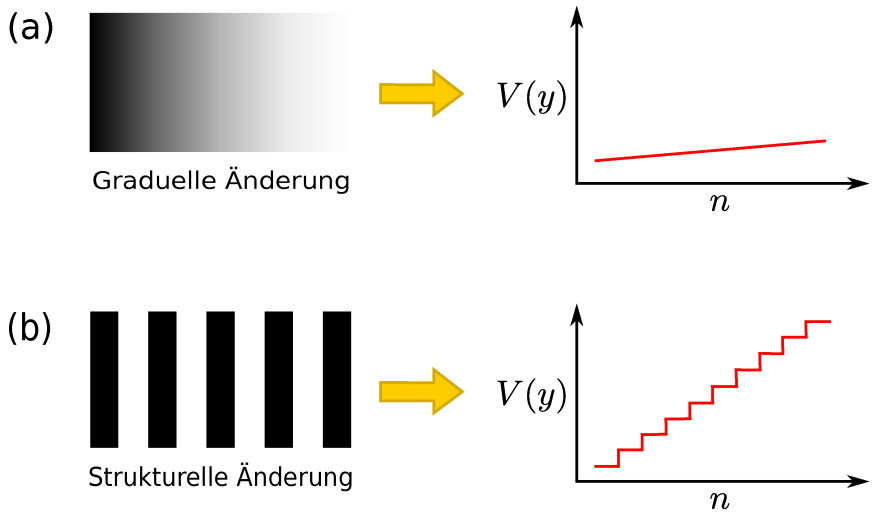

Abbildung 6.4.5: Totale Variation bei (a) gradueller und (b) struktureller Änderung des Signals.

einer Region ermittelt werden. Es sei $s(r)$ die Skalierungsstufe, aus der die Region r extrahiert wurde. Mit $F(s(r),i,j)$ wird dann der Merkmalswert des Bildpunktes an der Position $(i,j) \in r$ angegeben. Des Weiteren sei $|r|$ die Anzahl der Bildpunkte in r. Für das arithmetische Mittel gilt dann:

$$M(F,r) = \frac{1}{|r|} \sum_{(i,j)\in r} F(s(r),i,j). \tag{6.4.13}$$

Mit der Variation soll angegeben werden, wie stark sich die Merkmalswerte innerhalb eine Region ändern. Das Maß basiert auf dem Konzept der *totalen Variation* nach Rudin et al. [116]. Für ein digitales Signal, y, ist die totale Variation wie folgt definiert:

$$V(y) = \sum_n |y_{n+1} - y_n| \,. \tag{6.4.14}$$

Für zwei Dimensionen lässt sie sich zu

$$V(y) = \sum_{i,j} \sqrt{|y_{i+1,j} - y_{i,j}|^2 + |y_{i,j+1} - y_{i,j}|^2} \tag{6.4.15}$$

erweitern. Um die totale Variation einer Region relativ zu ihrer Größe anzugeben, wird diese Definition hier wie folgt angepasst:

$$V(F,r) = \frac{1}{|r|} \sum_{(i,j)\in r} \sqrt{|F(s(r),i+1,j) - F(s(r),i,j)|^2 + |F(s(r),i,j+1) - F(s(r),i,j)|^2}. \tag{6.4.16}$$

Die visuelle Distanz, D_r, wird schließlich wie folgt definiert:

$$D_r(F, r_k, r_i) = |M(F, r_k) - M(F, r_i)|^a + |V(F, r_k) - V(F, r_i)|^a. \quad (6.4.17)$$

Wählt man für a einen Wert größer als Eins, werden größere Unterschiede stärker betont. Ein Vorteil der totalen Variation ist, dass sie eine gewisse Robustheit gegenüber variierenden Lichtverhältnissen aufweist, wie in Abbildung 6.4.5 verdeutlicht wird. Graduelle Änderungen durch perspektivische Helligkeitsverläufe haben nur geringen Einfluss auf die totale Variation. Hingegen haben strukturelle Änderungen, die von texturellen Eigenschaften ausgehen, einen großen Einfluss auf die totale Variation.

6.4.3 *Merkmalsintegration*

Im nächsten Schritt, der Merkmalsintegration, werden die visuellen Distanzwerte der einzelnen Merkmalstypen zu einem einzigen Salienzwert zusammengefasst. Im Ergebnis erhält jede Region genau einen Salienzwert. Das Vorgehen hier basiert auf dem Modell von Itti et al. [57], wird aber auf die Betrachtung von Regionen angepasst. Die Menge der Merkmalstypen, $\mathscr{F}$, teilt sich in drei Kategorien auf, nämlich in Merkmale für Intensität, $\mathcal{I}$, Farbe, $\mathcal{C}$, und Orientierung, $\mathcal{G}$. Entsprechend der folgenden Gleichungen liefern alle drei Kategorien einen Beitrag zur Salienz:

$$\mathcal{I}(r_k) = D(I, r_k) \quad (6.4.18)$$

$$\mathcal{C}(r_k) = \frac{1}{2}\left(D(C_{RG}, r_k) + D(C_{BY}, r_k)\right) \quad (6.4.19)$$

$$\mathcal{G}(r_k) = \frac{1}{4}\left(D(G_{0^\circ}, r_k) + D(G_{45^\circ}, r_k) + D(G_{90^\circ}, r_k) + D(G_{-45^\circ}, r_k)\right) \quad (6.4.20)$$

Die Salienz für eine Region, r_k, wird nun zu

$$S(r_k) = \frac{1}{3}\left(\mathcal{I}(r_k) + \mathcal{C}(r_k) + \mathcal{G}(r_k)\right) \quad (6.4.21)$$

bestimmt.

Anhand der Salienzwerte können nun die salientesten Proto-Objekte zwecks Klassifizierung ausgewählt werden. Bei der Frage, wie viele Proto-Objekte weiterverarbeitet werden sollen, sind verschiedene Strategien denkbar. Steht beispielsweise eine bestimmte Bearbeitungszeit pro Bild zur Verfügung, kann man entsprechend die Anzahl an Regionen pro Bild auf eine feste Anzahl begrenzen. Denkbar ist auch die Wahl eines festen Schwellwerts. Es werden dann nur die Regionen weiterverarbeitet, deren Salienz diesen überschreiten. Auch eine Kombination aus beidem ist möglich. Zu beachten ist, dass die gewählten Regionen geometrisch überlappen können, wenn sie aus unterschiedlichen Skalierungsstufen

extrahiert wurden. Dies kann bedeuten, dass sich ein kleineres Objekt perspektivisch vor einem größeren befindet. Eine weitere Möglichkeit ist, dass die kleinere Region ein Teilobjekt eines größeren Objekts darstellt, welches wiederum durch die größere Region repräsentiert wird.

6.4.4 *Gerichtete Aufmerksamkeit*

Im Folgenden wird eine Variante, des im letzten Abschnitt vorgestellten Salienzdetektors präsentiert, bei der Modellwissen in den Detektionsprozess einbezogen wird. Hierfür wird eine Wissensbasis aufgebaut, die hier als Analogie zu dem visuell arbeitenden Gedächtnis des menschlichen Sehsystems verstanden wird, das in Abschnitt 4.1 betrachtet wurde. Die Bildung des Modellwissens erfolgt auf Basis von Trainingsbildern. Aus diesen werden die Salienzmerkmale von Instanzen der betrachteten Objektklassen sowie von Negativbeispielen aus Hintergrundbereichen extrahiert und gelabelt. Je nachdem worauf die Aufmerksamkeit gerichtet werden soll, ist sowohl die Betrachtung einer einzelnen Objektkategorie, als auch die Betrachtung mehrerer Objektkategorien möglich. Bei der Merkmalsextraktion wird stets die Umgebung einbezogen, in der das jeweilige Objekt auftritt. Die Salienzmerkmale des Objekts und seiner Umgebung werden in eine einheitliche Vektordarstellung überführt. Auf Basis dieser Vektoren wird ein Klassifizierer mittels eines überwachten Trainingsverfahrens angelernt. Es wird hierfür der Random-Forest-Klassifizierer eingesetzt, der in Abschnitt 2.3.2.1 beschrieben wurde. Der trainierte Klassifizierer wird schließlich für die Salienzdetektion eingesetzt. Abbildung 6.4.6 zeigt das Schema dieses Vorgehens. Die ersten Schritte bis zur Extraktion der regionalen Salienzmerkmale stimmen mit dem im letzten Abschnitt beschriebenen Bottom-Up-Verfahren überein. Die wesentliche Modifikation besteht in der Bildung von gelabelten Salienzvektoren, die für das überwachte Training verwendet werden. Um diese zu erstellen, wird zunächst auf allen Skalierungsstufen, l, für jede Region, $r_k \in \mathcal{R}_l$, durch Konkatenierung der Histogramme aller Merkmalstypen aus $\mathscr{F}$ ein Vektor, $\vec{h}_{r_k}$, gebildet:

$$\vec{h}_{r_k} = (H_{I,r_k}(1), ..., H_{I,r_k}(N), ..., H_{G_{-45^\circ},r_k}(1), ..., H_{G_{-45^\circ},r_k}(N)). \tag{6.4.22}$$

Die Größe der Vektoren entspricht $M = N|\mathscr{F}|$. Die Variante für Mittelwert und totale Variation lautet

$$\vec{h}_{r_k} = (M(I,r_k), V(I,r_k), ..., M(G_{-45^\circ},r_k), V(G_{-45^\circ},r_k)). \tag{6.4.23}$$

Hier gilt für die Größe $M = 2|\mathscr{F}|$. Aus den Vektoren wird schließlich für jede Region, $r_k \in \mathcal{R}_l$, ein Salienzmerkmalsvektor gebildet:

$$\vec{v}_{r_k} = (h_{r_k,1}, ..., h_{r_k,M}, b_{r_k,1}, ..., b_{r_k,M}). \tag{6.4.24}$$

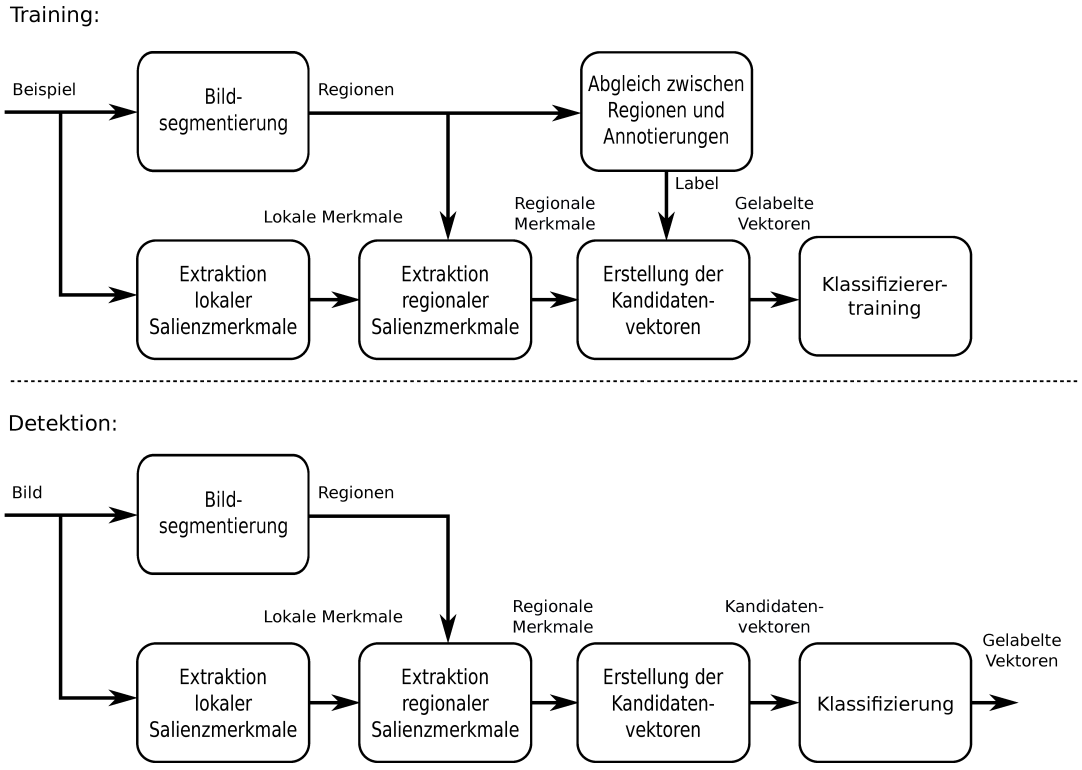

Abbildung 6.4.6: Salienzdetektion auf Basis von Modellwissen. Extraktion der Salienzmerkmale, Training und Klassifizierung.

Hierbei beschreibt der Vektor h_{r_k} die Region selbst und der Vektor b_{r_k} die Umgebung der Region. Letzterer wird nach folgender Gleichung gebildet:

$$\vec{b}_{r_k} = \frac{1}{|\mathcal{R}_l| - 1} \sum_{r_i \in \mathcal{R}_l \setminus r_k} \omega_s(r_i)\omega_d(r_k, r_i)\vec{h}_{r_i}. \tag{6.4.25}$$

Für das Labeln der Vektoren wird vorausgesetzt, dass die Trainingsbilder mit gelabelten Bounding-Boxen für die betrachtete Objektkategorie annotiert sind. Für jedes Segment wird überprüft, ob eine hinreichende Überschneidung mit einer Bounding-Box vorliegt. Hierfür wird das Kriterium

$$a_0 = \frac{\text{area}(B_s \cap B_a)}{\text{area}(B_s \cup B_a)} \tag{6.4.26}$$

nach Everingham et al. [117] verwendet, wobei B_s die Bounding-Box des Segments und B_a die annotierte Bounding-Box eines Positivbeispiels darstellt. Eine Überschneidung wird für $a_0 > 0.5$ angenommen. Liegt keine Überschneidung vor, wird der Merkmalsvektor als Negativbeispiel gelabelt. Die Detektion selbst verläuft parallel zum Trainingsverfahren. Der Klassifizierer erhält den Salienzmerkmalsvektor einer unbekannten Region als Eingabe und liefert als Ausgabe einen kontinuierlichen Wert zwischen Null und Eins. Je größer der Wert, desto

besser ist die Übereinstimmung mit der gesuchten Objektkategorie hinsichtlich der antrainierten Salienzmerkmale. Zu beachten ist, dass hierdurch lediglich eine Vorfilterung realisiert wird. Es wird noch nicht notwendiger Weise eine zuverlässige Aussage darüber geliefert, ob es sich tatsächlich um ein Objekt der gesuchten Kategorie handelt.

6.5 OBJEKTKLASSIFIZIERUNG

Nachdem die salienten Proto-Objekte detektiert wurden, werden sie nun im nächsten Schritt klassifiziert. Der hier verwendete Klassifizierer basiert auf dem Bag-Of-Features-Ansatz. Wie in Abschnitt 3.4 bereits ausgeführt wurde, bietet die Wahl dieses Ansatzes einige wichtige Vorteile. So liefern BoF-Modelle im Allgemeinen gute Ergebnisse und stellen dabei nur geringe Anforderungen an die verwendeten Trainingsdaten. Auch lässt sich der Ansatz problemlos auf eine größere Anzahl an Objektkategorien anwenden. Im Folgenden werden nun die Details der Klassifizierung beschrieben. Abschnitt 6.5.1 erläutert, wie auf Basis der Proto-Objekte die Merkmalsextraktion vorgenommen wird. Abschnitt 6.5.2 beschreibt das Vorgehen beim Trainingsverfahren und bei der Klassifizierung.

6.5.1 *Merkmalsextraktion*

Im Folgenden wird beschrieben, wie die BoF-Repräsentation eines Proto-Objekts erstellt wird. In Analogie zur menschlichen Wahrnehmung wird davon ausgegangen, dass das Proto-Objekt zum Zeitpunkt der Merkmalsextraktion fixiert ist, d.h. es befindet sich im fovealen Sichtbereich des Auges. Unabhängig davon auf welcher Skalierungsstufe das Proto-Objekt detektiert wurde, wird bei der Merkmalsextraktion davon ausgegangen, dass die Region des jeweiligen Proto-Objekts in der höchstmöglichen Auflösungsstufe zur Verfügung steht. Ausgehend von dieser wird das Objekt entsprechend Abbildung 6.5.1 extrahiert. Zunächst wird ein rechteckiger Bereich um die Region des Proto-Objekts ausgeschnitten (b). Dieser wird anschließend bezüglich der Auflösung normalisiert (c). Zur Festlegung des

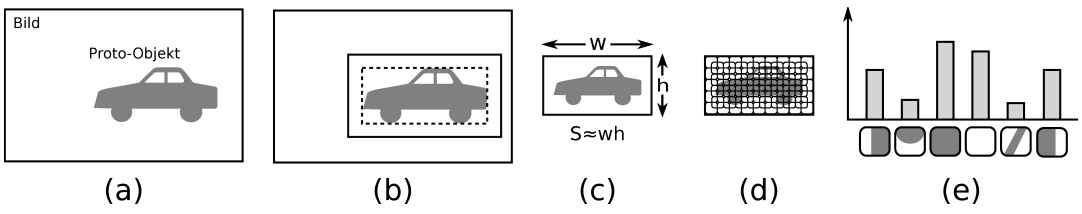

Abbildung 6.5.1: Merkmalsextraktion. (a) Detektiertes Proto-Objekt. (b) Extraktion der Region. (c) Normalisierung der Größe. (d) Extraktion der lokalen Merkmalsvektoren. (e) Erstellen des globalen Merkmalsvektors.

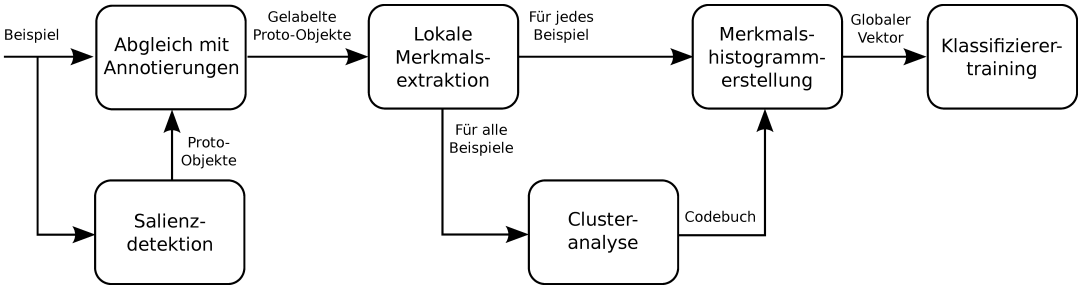

Abbildung 6.5.2: Schematische Darstellung des Trainingsverfahrens.

Bereichs wird die Bounding-Box des Proto-Objekts derart vergrößert, dass bei normalisierter Auflösung zu allen Seiten eine halbe Kachelbreite zusätzlich zur Verfügung steht. Dadurch wird erreicht, dass die Kontur des Objekts bei der Merkmalsextraktion erfasst werden kann. Des Weiteren kann der unmittelbare Bereich, in dem ein Objekt auftritt, angelernt werden. Die Annahme dabei ist, dass sich dies positiv auf die Klassifizierungsergebnisse auswirkt. Im nächsten Schritt folgt die Extraktion der lokalen Merkmale (d). Diese werden nach dem DSIFT-Verfahren (siehe Abschnitt 3.3) gleichmäßig über dem Bild verteilt extrahiert. Liegen Teile des Rechteckbereiches außerhalb des Bildes, werden diese bei der Extraktion der lokalen Merkmale nicht berücksichtigt. Die lokalen Merkmale werden schließlich nach dem Bag-Of-Features-Prinzip durch Vektorquantisierung und Histogrammbildung (siehe Abschnitt 2.3.3) zu einem globalen Merkmalsvektor zusammengefasst, der dann das Proto-Objekt repräsentiert.

6.5.2 *Training und Klassifizierung*

Im Trainingsverfahren wird zum einen das Codebuch erstellt, welches für die Vektorquantisierung benötigt wird. Hierfür wird das k-Means-Verfahren verwendet, das in Abschnitt 2.3.3 erläutert wurde. Zum anderen wird der Klassifizierer trainiert. Es wird der Random-Forest-Klassifizierer eingesetzt, der in Abschnitt 2.3.2.1 beschrieben wurde. Abbildung 6.5.2 zeigt eine schematische Darstellung des Trainingsverfahrens. Im ersten Schritt durchläuft das Trainingsbild den Salienzdetektor. Durch das Vorschalten der Salienzdetektion wird erreicht, dass nur solche Regionen in das Training einfließen, die auch beim Detektionsverfahren den Salienzdetektor passieren würden. Die so detektierten Proto-Objekte werden im nächsten Schritt mit den Annotationen des Trainingsbildes abgeglichen. Es wird hierfür vorausgesetzt, dass die Trainingsbilder mit gelabelten Bounding-Boxen für alle betrachteten Objektkategorien versehen sind. Für jedes Proto-Objekt wird überprüft, ob eine hinreichende Überschneidung mit einer Bounding-Box vorliegt. Hierfür wird wieder das Kriterium nach Russell et al. [10] verwendet (siehe Abschnitt 6.4.4). Liegen keine hinreichenden Überschneidungen vor, wird das Proto-Objekt als negatives Trainingsbeispiel gelabelt. Die nachfolgende

Extraktion der lokalen Merkmale sowie das Erstellen der Histogramme erfolgt entsprechend den Ausführungen des letzten Abschnitts. Zusätzlich werden zur Erstellung des Codebuchs die lokalen Merkmalsvektoren aller Trainingsbeispiele zunächst gesammelt und anschließend geclustert.

Das Klassifizierungsverfahren läuft analog zum Training ab und besteht aus den bereits bekannten Schritten. Für jedes detektierte Proto-Objekt wird der globale Merkmalsvektor extrahiert und anschließend klassifiziert. Der Random-Forest-Klassifizierer liefert als Ausgabe für jede Kategorie einen kontinuierlichen Zuversichtswert. Durch die Festlegung eines Schwellwertes können hieraus Label erzeugt werden. Die Höhe des Wertes hat dabei einen direkten Einfluss auf die Detektions- und Fehlerraten.

7 EVALUIERUNG

Im Folgenden wird die Evaluierung der im letzten Kapitel vorgestellten Methoden erläutert. In Abschnitt 7.1 werden zunächst verschiedene öffentlich zugängliche Bilddatenbanken aufgezeigt und deren Vor- und Nachteile erörtert. Anschließend werden in Abschnitt 7.2 die verwendeten Evaluierungsmethoden beschrieben. Abschnitt 7.3 zeigt eine Übersicht der Evaluierungsergebnisse. In Abschnitt 7.4 werden Vergleiche mit verschiedenen Verfahren aus der Literatur gezogen. Schließlich werden in Abschnitt 7.5 einige Aspekte der Evaluierungsergebnisse näher im Detail betrachtet. Dabei wird unter anderem untersucht, wie sich unterschiedliche Objekteigenschaften wie Größe und Perspektive auf die Ergebnisse auswirken.

7.1 BILDDATENBANKEN

Zur Durchführung der Evaluierung wird eine nicht geringe Anzahl an Beispielbildern benötigt, die über geeignete Annotationen verfügen müssen. Eine Möglichkeit der Annotierung besteht in der Verwendung eines Eye-Tracking-Systems. Dabei werden die sakkadischen Augenbewegungen verschiedener Personen bei der Betrachtung von Bildern aufgezeichnet, um daraus Fixationspunkte abzuleiten. Solche Datenbanken werden typischerweise verwendet, wenn ein Aufmerksamkeitsmodell eingesetzt werden soll, um menschliche Augenbewegungen vorherzusagen oder zu simulieren. Unter den öffentlich zur Verfügung stehenden Datenbanken sind hier vor allem MIT300 [118] und CAT2000 [119] zu nennen, da sie einen vergleichsweise großen Umfang aufweisen. So verfügt MIT300 über 300 Bilder mit Aufzeichnungen von 39 Betrachtern im Alter zwischen 18 und 50 Jahren und CAT2000 über 2000 Bilder mit Aufzeichnungen von 24 Betrachtern im Alter zwischen 18 und 27 Jahren.

Bei der hiesigen Betrachtung stehen Augenbewegungen allerdings nicht im Fokus. Vielmehr sind Annotationen bezüglich der dargestellten Objekte von Bedeutung. Solche Annotationen stehen bei den oben genannten Datenbanken jedoch nicht zur Verfügung. Eine Alternative bietet hier die MSRA-Salient-Object-Database [120]. Wie in Abbildung 7.1.1 zu sehen ist, handelt es sich um Bilder, die jeweils ein salientes Objekt zeigen, das mit einer Bounding-Box annotiert ist. Die Datenbank besteht aus zwei Datensätzen. Der erste Satz besteht aus 20.000 Bildern, die von drei Bearbeitern annotiert wurden. Der zweite Satz besteht aus 5.000 Bildern, die von neun Bearbeitern annotiert wurden. Label zu den annotierten Objekten sind nicht vorhanden, sodass die Datensätze

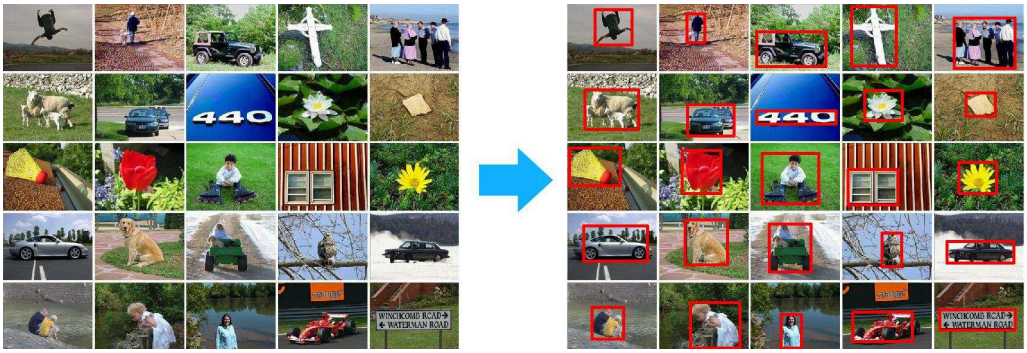

Abbildung 7.1.1: Auszug aus der MSRA-Salient-Object-Database [120].

nicht für Klassifizierungsprobleme verwendet werden können. Problematisch an der Datenbank ist auch, dass die Problemstellung der Salienzdetektion durch die Auswahl der Bilder stark idealisiert wird, da es stets gilt, einen einzelnen und häufig zentralen Vordergrundbereich vor einem meist relativ eintönigen Hintergrund zu lokalisieren. Die Aufgabenstellung ließe sich somit auch als Zwei-Klassen-Segmentierungsproblem formulieren. Häufig stellt sich das Problem der Lokalisierung interessanter Objekte nicht so eindeutig dar, wenn sich die Bildinhalte unübersichtlicher gestalten. Mehr Vielfalt bieten hier Datenbanken, die speziell auf das Problem der Objektlokalisierung in komplexen Szenen ausgelegt sind. Einen recht großen Umfang bieten hier die PASCAL-VOC-Datenbanken, die im Rahmen der *PASCAL-Visual-Object-Classes-Challenge* [117] zwischen den Jahren 2005 und 2012 jährlich in stetig überarbeiteten Versionen herausgegeben wurden. Die Aufgabenstellungen umfassen unter anderem die Lokalisierung und Klassifizierung von Objekten. Für alle Aufgabenstellungen sind Benchmarks mit einheitlichen Evaluierungsmethoden definiert worden (vgl. Abschnitt 7.2). Diese haben in der Fachwelt eine breite Akzeptanz, sodass Ergebnisse für viele Verfahren aus der Literatur veröffentlicht worden sind. In dieser Arbeit werden die Versionen VOC2007 [121] und VOC2012 [122] betrachtet. Die VOC2007-Datenbank ist nach wie vor von Bedeutung, da für sie viele Vergleichsergebnisse aus der Literatur zur Verfügung stehen. Sie enthält 5011 Bilder mit insgesamt 12608 annotierten Objekten aus zwanzig unterschiedlichen Objektkategorien[1]. Die VOC2012-Datenbank verfügt über die gleichen Objektkategorien. Sie hat den Vorteil, dass sie über einen größeren Umfang von 11540 Bildern mit insgesamt 27450 annotierten Objekten verfügt. Abbildung 7.1.2 zeigt einen Auszug hieraus. Wie die Beispiele zeigen, kommen sowohl Bilder mit einem einzelnen, großen, zentralen Objekt als auch Bilder mit mehreren kleinen bis mittelgroßen Objekten an verschiedenen Bildpositionen vor. Die Annotationen enthalten neben der Bounding-Box und einem Label auch noch eine grobe Angabe zur Pose (frontal,

1 Dies sind im Einzelnen: Flugzeug, Fahrrad, Vogel, Boot, Flasche, Bus, Auto, Katze, Stuhl, Kuh, Esstisch, Hund, Pferd, Motorrad, Person, Topfpflanze, Schaf, Sofa, Zug und Bildschirm.

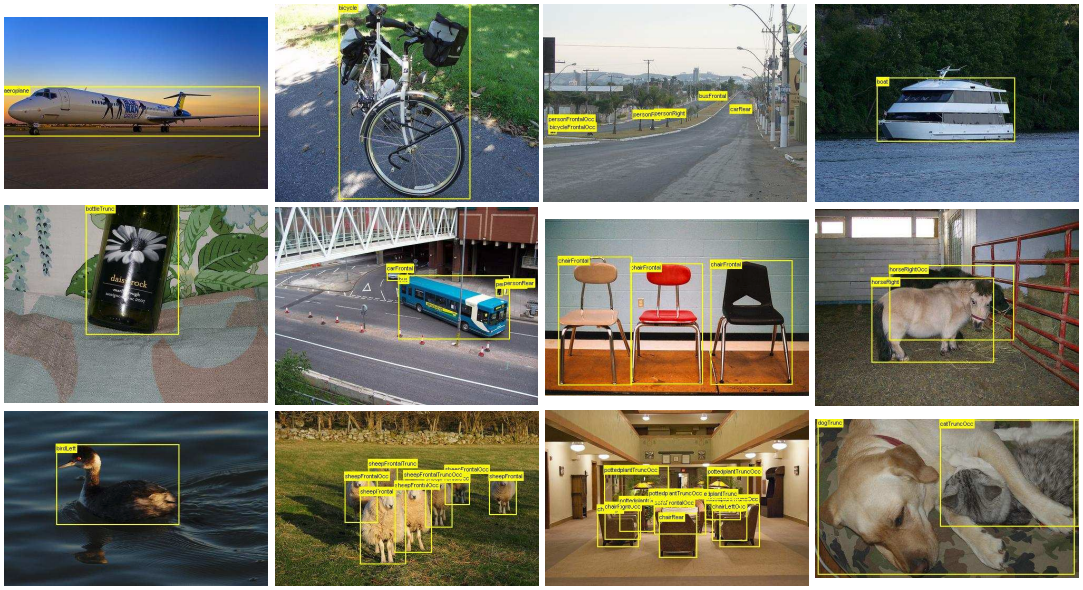

Abbildung 7.1.2: Auszug aus der VOC2012-Datenbank [122].

links, rechts oder rückseitig) und einen Hinweis darauf, ob das Objekt teilweise verdeckt (Occ) oder am Bildrand abgeschnitten (Trunc) ist. Darüber hinaus sind einige Objekte mit dem Hinweis versehen, das diese besonders schwer zu erkennen sind. Die Datenbank ist in einen Trainings- und einen Validierungsdatensatz aufgeteilt, die in etwa gleich groß sind.

In dieser Arbeit geht es um die Detektion interessanter Objekte, auf die ein Betrachter typischerweise aufmerksam wird. Bei der Verwendung einer Datenbank für die Objektklassifizierung muss dabei bedacht werden, dass es vorkommen kann, dass ein detektierter Bildbereich zwar ein interessantes Objekt zeigt, dieses jedoch nicht annotiert wurde. Dies wird in der Regel daran liegen, dass Datenbanken für bestimmte Aufgabenstellungen erstellt werden und dann entsprechend nur solche Objekte annotiert werden, die für dieses Aufgabenstellung auch von Bedeutung sind. Aus dem gleichen Grund kann es annotierte Objekte geben, die eher nicht interessant sind, da sie im Vergleich zu anderen Bildinhalten nur eine untergeordnete Rolle spielen. Elazary und Itti [3] haben deshalb für ihre Experimente Bilder der *LabelMe* Datenbank verwendet, da bei dieser Objekte ohne eine bestimmte Aufgabenstellung annotiert wurden. Sie argumentieren, dass ein Bearbeiter, der beliebige Objekte annotieren soll, vornehmlich solche Objekte auswählen wird, die er interessant findet. Schließlich wird er bei der Bearbeitung naturgemäß von seinen eigenen Aufmerksamkeitsprozessen beeinflusst. Jedoch wird auch darauf hingewiesen, dass es noch andere Faktoren geben mag, die einen Bearbeiter dazu bewegen können, bestimmte Bildinhalte zu bevorzugen. Im Falle der PASCAL-VOC-Datenbanken richten sich die Annotationen hingegen nach den betrachteten Klassen. Es stellt sich also die Frage, in wie weit es sich bei den annotierten Objekten auch um die interessanten Objekte

handelt. Dies muss nicht unbedingt überall gegeben sein. Jedoch spielt auch hier der menschliche Faktor eine große Rolle, da die Bilder von Menschen aufgenommen, zusammengestellt und annotiert wurden. Es wird hier deshalb davon ausgegangen, dass zumindest eine starke Korrelation zwischen den interessanten und den annotierten Objekten besteht.

Als weiterer Punkt muss bei der Verwendung der PASCAL-VOC-Datenbanken beachtet werden, dass sie sich hauptsächlich aus Bildern zusammensetzen, die vom jeweiligen Fotografen aufgenommen wurden, um bestimmte Inhalte einzufangen. Das bedeutet, dass Objekte häufig im Fokus der Aufnahme stehen, hohe Kontrastwerte aufweisen und sich in der Nähe des Bildmittelpunkts befinden. Meistens wurden die Bilder zudem bei guten Lichtverhältnissen aufgenommen. Hierbei handelt es sich um Eigenschaften, die typischerweise bei Top-Down-Ansätzen zur Verbesserung der Ergebnisse ausgenutzt werden.

7.2 METHODEN

Das im letzten Kapitel präsentierte Verfahren muss zur erfolgreichen Detektion eines interessanten Objekts nacheinander drei Teilprobleme lösen. Als Erstes muss das Objekt mit ausreichender Genauigkeit segmentiert werden, damit es in den Kandidatenpool der Proto-Objekte aufgenommen wird. Zweitens muss das Proto-Objekt als salient eingestuft werden. Als Drittes muss das saliente Proto-Objekt schließlich auch richtig klassifiziert werden. Bei der Evaluierung werden diese drei Teilprobleme nacheinander betrachtet. Die dabei verwendeten Methoden werden im Folgenden erläutert.

Bei der Segmentierung gilt es, mit einem möglichst kleinen Kandidatenpool eine möglichst hohe Abdeckung der annotierten Objekte zu erzielen. Die Idee der multiskalaren Segmentierung besteht darin, den Anteil dieser Objekte zu erhöhen, was wiederum auf Kosten eines größeren Kandidatenpools geht. Die Bewertung der Genauigkeit erfolgt auf Basis der Bounding-Box eines Segments. Als Genauigkeitskriterium wird die bereits im letzten Kapitel aufgeführte Gleichung der PASCAL-VOC-Challenge [117] verwendet:

$$a_0 = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})}. \tag{7.2.1}$$

Für einen gültigen Treffer wird $a_0 > 0.5$ gefordert. Das Vorgehen ist nun wie folgt. Das Segmentierungsverfahren wird zunächst auf das Eingabebild angewendet. Von allen Segmenten werden dann die zugehörigen Bounding-Boxen bestimmt. Für alle annotierten Objekte wird dann die jeweils am besten übereinstimmende Bounding-Box ermittelt. Dabei wird der Grad der Übereinstimmung anhand des Kriteriums a_0 bewertet. Es wird also jeweils die Bounding-Box gesucht, die a_0 maximiert. Neben den Werten für a_0 wird des Weiteren noch die Gesamtzahl der Segmente bestimmt, die für das Eingabebild ermittelt wurden. Das Vorgehen

wird für alle Bilder des Evaluierungsdatensatzes wiederholt. Aus den Ergebnissen wird schließlich der Gesamtanteil an annotierten Objekten ermittelt, die mit hinreichender Genauigkeit segmentiert wurden. Dieser wird der durchschnittlichen Anzahl an Segmenten pro Bild gegenübergestellt.

Die Bewertung der Salienzdetektion erfolgt analog zu Elazary und Itti [3], die den Anteil der detektierten Objekte der Anzahl an Aufmerksamkeitsverschiebungen gegenüberstellen. Es wird hier wie folgt vorgegangen. Zunächst wird das Salienzdetektionsverfahren auf das Eingabebild angewendet. Anschließend wird in nicht-aufsteigender Reihenfolge der Salienzwerte über die Regionen iteriert. Dies entspricht der Winner-Take-All-Strategie, bei der sich die Aufmerksamkeit von Region zu Region verschiebt. Der aktuellen Region wird zunächst ein Rang zugewiesen, der der Anzahl an bisherigen Aufmerksamkeitsverschiebungen entspricht. Wird eine Region bei der n-ten Iteration aufgesucht, hat sie entsprechend den Rang n. Anschließend wird die Region mit den annotierten Objekten abgeglichen. Hierbei wird wieder das Kriterium $a_0 > 0.5$ angewendet. Gibt es einen Treffer, gilt das entsprechende Objekt als detektiert. Gibt es für ein Objekt mehrere Treffer, wird nur der niedrigste Rang berücksichtigt. Das Vorgehen wird für den gesamten Evaluierungsdatensatz wiederholt. Aus den Ergebnissen wird für jeden Rang der Gesamtanteil an Objekten ermittelt, der bis zu diesem Rang detektiert wurde.

Für die Evaluierung der Klassifizierungsergebnisse werden die Methoden der PASCAL-VOC-Challenge [122] angewendet. Diese basieren auf der Auswertung der *Genauigkeit* (engl.: precision) und der *Trefferquote* (engl.: recall). Die Genauigkeit gibt die richtigen positiven Befunde im Verhältnis zu allen positiven Befunden an, während die Trefferquote die richtigen positiven Befunden ins Verhältnis zu allen positiven Beispielen setzt.

$$\text{Genauigkeit} = \frac{\text{richtige Positive}}{\text{richtige Positive} + \text{falsche Positive}} \tag{7.2.2}$$

$$\text{Trefferquote} = \frac{\text{richtige Positive}}{\text{richtige Positive} + \text{falsche Negative}} \tag{7.2.3}$$

Liefert der Klassifizierer für jeden Befund einen Zuversichtswert, lassen sich daraus für verschiedenen Genauigkeiten die zugehörigen Trefferquoten ermitteln und in einer entsprechenden Kurve darstellen. Um unterschiedliche Verfahren einfacher miteinander vergleichen zu können, empfiehlt es sich, die *gemittelte Genauigkeit* (engl.: average precision) anzugeben. Hierbei werden die Genauigkeiten für vorgegebene Trefferquoten (i.d.R. im Intervall $[0, 0.1, ..., 1]$) gemittelt:

$$\text{AP} = \frac{1}{11} \sum_{r \in \{0, 0.1, ..., 1\}} p_i(r). \tag{7.2.4}$$

Hierbei ist $p_i(r)$ die Genauigkeit zur Trefferquote r. Es handelt sich allerdings um eine Näherung zur tatsächlichen Genauigkeit, $p(r)$:

$$p_i(r) = \max_{\tilde{r}:\tilde{r}\geq r} p(\tilde{r}). \tag{7.2.5}$$

Da die vorgegebenen Trefferquoten im Allgemeinen nicht exakt getroffen werden, wird die Genauigkeit maximiert, ohne dabei die jeweils vorgegebene Trefferquote zu unterschreiten.

7.3 ERGEBNISÜBERSICHT

Aufbauend auf den im letzten Abschnitt beschriebenen Evaluierungsmethoden wird nun im Folgenden eine Übersicht zu den Ergebnissen für die im letzten Kapitel vorgestellten Methoden präsentiert. Als Erstes sollen hier die Ergebnisse der Segmentierung betrachtet werden. Diese spielen für die Stärke des Gesamtverfahrens eine wichtige Rolle. Tabelle 7.3.1 zeigt die Ergebnisse für die unterschiedlichen Objektkategorien auf dem VOC2012-Validierungsdatensatz. Zusätzlich ist das Ergebnis über alle Objekte sowie eine Mittelung über die Objektkategorien angegeben, wobei jede Klasse gleichstark gewichtet wurde. Eingetragen ist jeweils die Anzahl an Objekten, die durchschnittliche Segmentierungsgenauigkeit, $\overline{a_0}$, sowie der Anteil der Objekte, bei denen mindestens ein Segment das Genauigkeitskriterium erfüllt. Die durchschnittliche Anzahl an Segmenten pro Bild beträgt hier 95.613 ± 14.662. Bei der dynamischen Anpassung des Schwellwerts wurde lediglich ein grober Rahmen gesetzt. Bei diesem wird angenommen, dass bei mehr als einhundert Regionen auf der ersten Skalierungsstufe eine Über- und bei weniger als 25 Regionen eine Untersegmentierung vorliegt, was eine Anpassung des Schwellwerts zur Folge hat. Auf eine Optimierung dieser Grenzen für das betrachtete Benchmark wurde im Sinne der verfolgten Bottom-Up-Strategie verzichtet. Bei den Ergebnissen kommt es zwischen den einzelnen Objektkategorien zu starken Unterschieden. So befindet sich die Kategorie *Person* relativ weit unten in der Liste. Problematisch ist hier, dass Personen bei der Segmentierung häufig in mehrere Bereiche zerlegt werden. Da der Konturübergang beispielsweise zwischen Kopf und Oberbekleidung meist stärker ausgeprägt ist als zwischen Kopf und Umgebung, führt auch die hierarchische Segmentierungsstrategie häufig nicht zum Erfolg. Ohne spezifisches Modellwissen sind solche Probleme im Allgemeinen recht schwierig zu adressieren. Einige Bildbeispiele hierzu folgen später in Abschnitt 7.5.

Im Folgenden sollen nun die Ergebnisse der Salienzdetektion betrachtet werden. Diese sind in Abbildung 7.3.1 für den VOC2012-Validierungsdatensatz dargestellt. Eingetragen sind die Kurven für die drei unterschiedlichen visuellen Distanzmaße aus Abschnitt 6.4.2 sowie für die gerichtete Aufmerksamkeit nach dem in Abschnitt 6.4.4 vorgeschlagenen Verfahren. Um den grundsätzlichen Nutzen der Salienzauswertung zu demonstrieren, wurde eine weitere Kurve

Kategorie	Anzahl	$\overline{a_0}$	$a_0 > 0.5$
Katze	618	0.705 ± 0.174	86.08%
Sofa	387	0.718 ± 0.195	85.53%
Hund	773	0.649 ± 0.184	80.08%
Bildschirm	414	0.677 ± 0.204	79.95%
Bus	320	0.610 ± 0.193	73.75%
Zug	329	0.613 ± 0.189	72.34%
Esstisch	374	0.610 ± 0.193	71.66%
Motorrad	376	0.599 ± 0.197	71.01%
Pferd	373	0.599 ± 0.190	70.24%
Kuh	347	0.583 ± 0.230	66.28%
Fahrrad	380	0.579 ± 0.203	64.47%
Kategorien gemittelt	**20**	$\mathbf{0.569 \pm 0.086}$	**63.87%**
Flugzeug	484	0.571 ± 0.237	63.84%
Alle Objekte	**15787**	$\mathbf{0.538 \pm 0.234}$	**58.28%**
Schaf	485	0.522 ± 0.260	57.53%
Stuhl	1449	0.549 ± 0.221	57.21%
Person	5110	0.498 ± 0.223	50.96%
Vogel	629	0.477 ± 0.243	50.56%
Auto	1173	0.480 ± 0.249	49.36%
Boot	491	0.467 ± 0.231	48.27%
Topfpflanze	542	0.469 ± 0.251	45.39%
Flasche	733	0.397 ± 0.240	32.88%

Tabelle 7.3.1: Segmentierungsergebnisse für die unterschiedlichen Objektkategorien auf dem VOC2012-Validierungsdatensatz.

eingetragen, bei der die Segmente des Kandidatenpools nach einer zufälligen Reihenfolge ausgewählt worden sind. Bei den Histogrammen wurden stets 16 Kanäle verwendet. Für das komplexere Histogrammdistanzmaß nach Gleichung 6.4.11 wird im Folgenden das Kürzel H1 verwendet. Das einfachere Histogrammdistanzmaß nach Gleichung 6.4.12 wird mit dem Kürzel H2 belegt. Mit MV wird die Variante auf Basis des arithmetischen Mittelwerts und der totalen Variation nach Gleichung 6.4.17 bezeichnet. Für das Verfahren der gerichteten Aufmerksamkeit nach Abschnitt 6.4.4 wird das Kürzel GA verwendet. Bei diesem wurden alle zwanzig Objektklassen der VOC2012-Datenbank gleichzeitig betrachtet, sodass nur zwischen den Klassen *Objekt* und *Nicht-Objekt* unterschieden wird. Die entsprechenden Merkmalsvektoren für das Training wurden aus

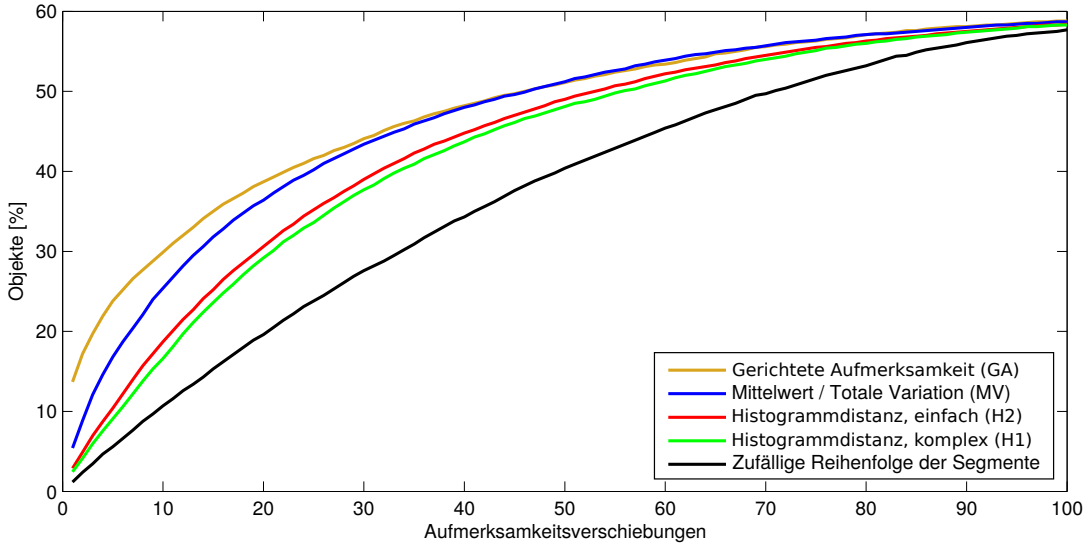

Abbildung 7.3.1: Ergebnisse der Salienzdetektion auf dem VOC2012-Validierungsdatensatz.

dem VOC2012-Trainingsdatensatz extrahiert. Ein Merkmalsvektor besteht dabei aus 28 Werten, die sich aus arithmetischem Mittelwert und totaler Variation der sieben Salienzmerkmale jeweils für die Region selbst und deren Umgebung zusammensetzen. Die Ergebnisse zeigen, dass MV als Distanzmaß besser geeignet ist als die histogrammbasierten Distanzmaße, H1 und H2. Bei letzteren erzielt H2 etwas bessere Ergebnisse. Alle salienzbasierten Ansätze haben gegenüber dem rein auf Segmentierung basierenden Ansatz einen deutlichen Vorsprung. Bei GA ist durch die gelernten Informationen vor allem eine Verbesserung der Ergebnisse im vorderen Bereich der Kurve zu beobachten. Abbildung 7.3.2 zeigt die Ergebnisse aufgeschlüsselt für die einzelnen Objektkategorien. Auch hier kommt es wie bei der Segmentierung zu recht großen Unterschieden zwischen den einzelnen Objektkategorien. Dies liegt im Wesentlichen daran, dass die Detektionsergebnisse auf den Segmentierungsergebnissen aufbauen. Allen Kurven gemein ist, dass sie zunächst schnell ansteigen und anschließend in die Sättigung übergehen. Bei GA konnte bei den meisten Objektkategorien im vorderen Bereich im Vergleich zu MV eine deutliche Verbesserung erzielt werden.

Ein erstes Bildbeispiel wird in Abbildung 7.3.3 dargestellt. Um die Funktionsweise der Methoden zu veranschaulichen, wurde hier ein recht einfaches Beispiel gewählt, bei dem sich ein großes Objekt vor einem monotonen Hintergrund mit graduellem Helligkeitsverlauf befindet. Reihe (a) zeigt die Gauß-Pyramide des Bildes und Reihe (b) die zugehörigen Segmente. Zur Darstellung der Segmente wird der über das jeweilige Segment gemittelte Farbwert verwendet. Die Ränder der Segmente sind etwas hervorgehoben, um die Grenzen zwischen den Segmenten besser sichtbar zu machen. Die Reihen (c), (d), (e) und (f) zeigen die Ergebnisse der Salienzdetektion für die unterschiedlichen visuellen Distanzmaße

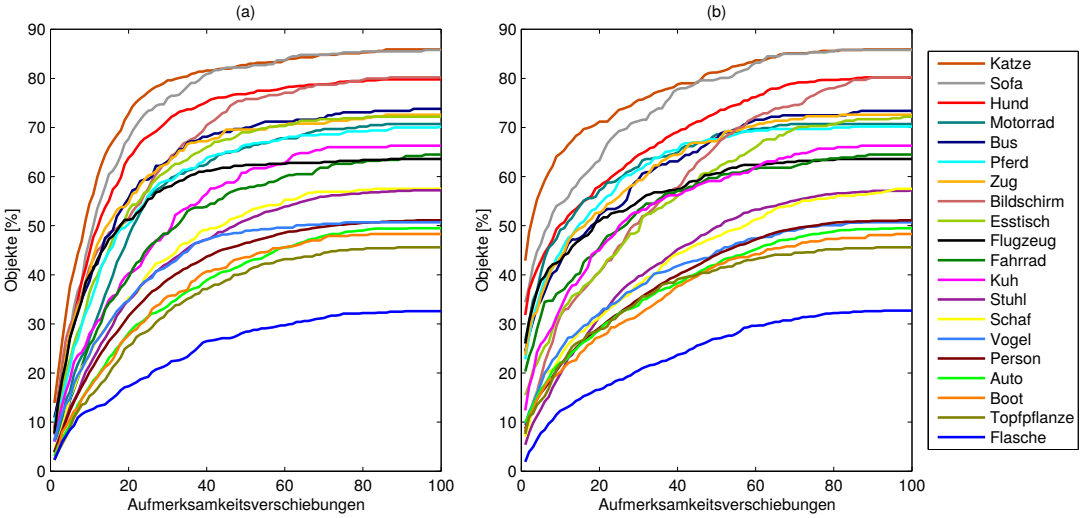

Abbildung 7.3.2: Ergebnisse der Salienzdetektion auf dem VOC2012-Validierungsdatensatz für die einzelnen Objektkategorien: (a) MV. (b) GA.

(MV, H2 und H1) und für die Variante der gerichteten Aufmerksamkeit (GA). Bei der Darstellung ist der Graustufenwert einer Region proportional zu seiner Salienz. Die salienteste Region der Pyramide ist weiß, die am wenigsten saliente Region schwarz dargestellt. Wie das Beispiel zeigt, stellt das Segmentierungsverfahren aufgrund der dynamischen Schwellwertanpassung ausgewogen über alle Skalierungsstufen Objekthypothesen auf. Da es sich bei dem Flugzeug um ein großes Objekt handelt, wird es erst in der vierten Stufe getroffen. Wird in der zweiten Stufe noch die hellste Region links unten im Bild als salient eingestuft, hebt sich bei Betrachtung des Flugzeugs als Ganzes in der vierten Stufe dieses vom flächenmäßig größeren Hintergrund ab. Bei MV ist dies deutlicher ausgeprägt als bei den Histogrammdistanzmaßen, H1 und H2, da bei letzteren der Helligkeitsverlauf stärker einfließt und die verschiedenen Himmelsregionen als weniger ähnlich wahrgenommen werden. Obwohl sich ein strahlend blauer Himmel in der Regel als salient darstellt, tritt er bei GA in den Hintergrund, da hier gelernt wurde, dass entsprechende Regionen eher keine Objekte zeigen. Entsprechend setzt sich hier das Flugzeug am deutlichsten ab.

Im Folgenden werden nun die Ergebnisse der Objektklassifizierung präsentiert. Das Verfahren wurde in Abschnitt 6.5 beschrieben. Die Identifizierung der interessanten Objekte ist ein wichtiger Aspekte zum Verständnis einer Szene. Wie eingangs in Kapitel 1 erläutert wurde, wird hier der Ansatz favorisiert, nur sehr wenige Regionen auszuwerten, um sich schnell und effizient einen Eindruck der Szene verschaffen zu können. Tabelle 7.3.2 zeigt die Ergebnisse des Präsenzproblems des VOC2012-Benchmarks für die Auswertung der zehn salientesten Regionen. Dabei wird eine BoF-Repräsentation über alle zehn Regionen erstellt und ausgewertet. Beim k-Means-Clustering wird für k der Wert 1024 gewählt.

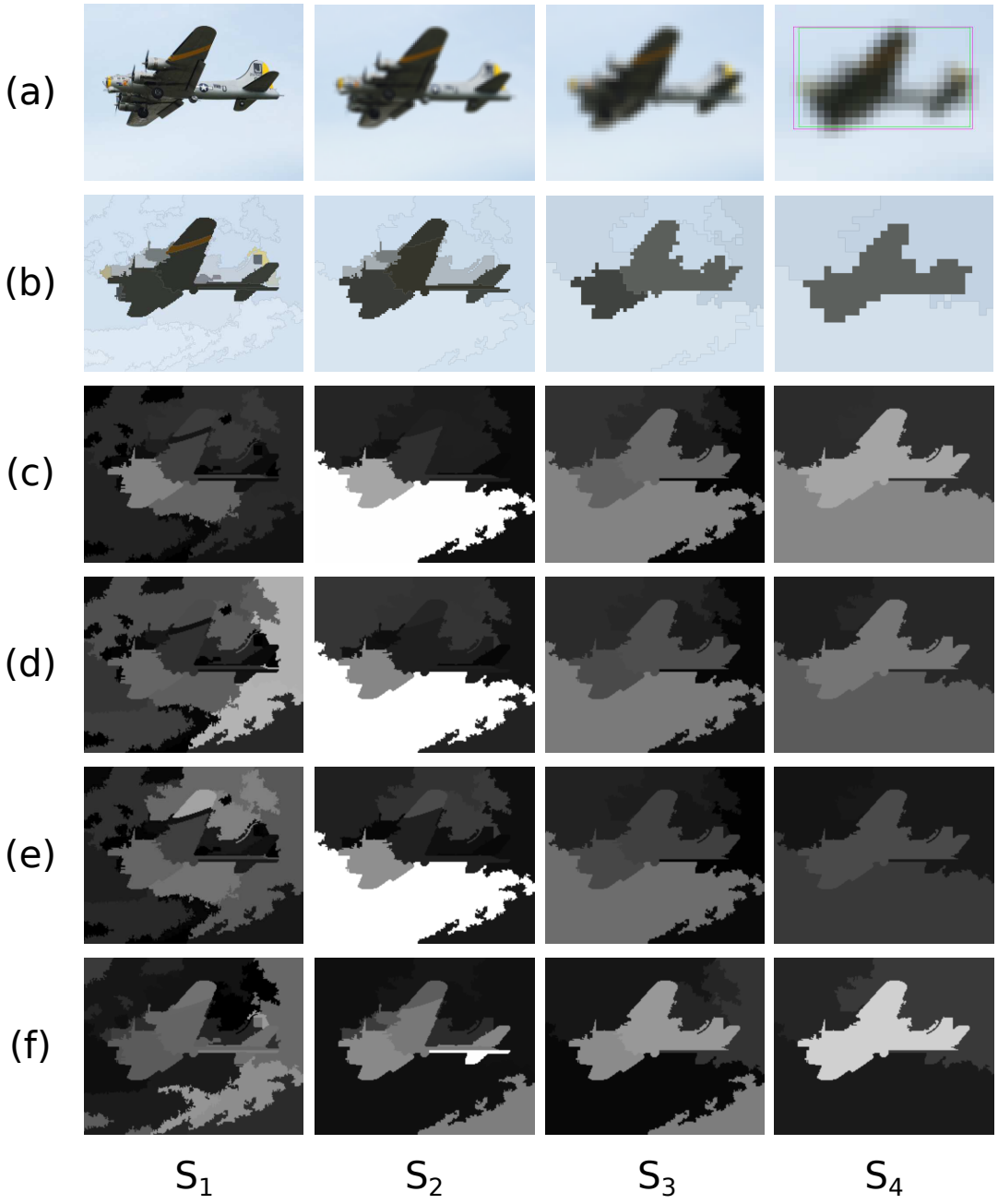

Abbildung 7.3.3: Bildbeispiel. (a) Gauß-Pyramide. (b) Segmentierung. (c) MV. (d) H2. (e) H1. (f) GA.

Das bedeutet, es wird ein entsprechendes Codebuch mit 1024 Einträgen verwendet. Die SIFT-Merkmale werden auf einem dichten Raster im Abstand von einem Pixel extrahiert. Für jede Objektklasse wird auf dem Trainingsdatensatz ein Random-Forest-Klassifizierer mit jeweils 10.000 Entscheidungsbäumen antrainiert. In der Tabelle sind jeweils die AP-Werte für die unterschiedlichen Objektkategorien sowie der AP-Wert gemittelt über alle Kategorien eingetragen.

Letzterer beträgt 30.0%. Auf der anderen Seite des Spektrums stehen Verfahren, die eine sehr intensive Auswertung der Trainingsdaten und Eingabebilder vornehmen. Für die entsprechenden Verfahren, die bei der PASCAL-VOC-Challenge 2012 [122] bewertet wurden, liegen die Ergebnisse für den gemittelten AP-Wert zwischen 36.4% und 82.2%.

7.4 LITERATURVERGLEICHE

Im Folgenden werden verschiedene Bottom-Up-Verfahren aus der Literatur mit dem in dieser Arbeit vorgeschlagenen Bottom-Up-Ansatz verglichen. Betrachtet wird hier die Variante MV, die die visuelle Distanz mittels arithmetischem Mittel und totaler Variation bewertet. Die Verfahren, die hier verglichen werden, sind in Kapitel 5 bereits betrachtete worden. Das Verfahren von Zhai und Shah [87] basiert auf Histogrammen über Farbwerte. Hou und Zhang [84] ermitteln die Salienz mit Hilfe des spektralen Rests. Bei Achanta et al. [86] wird Salienz auf Basis von Bandpassfiltern ermittelt. Cheng et al. [96] ermitteln Salienz über die Differenz von Farbhistogrammen, die über Segmente gebildet werden. Abbildung 7.4.1 zeigt die Ergebnisse für den Validierungsdatensatz des VOC2012-Benchmarks. Als Orientierungshilfe ist eine weitere Kurve eingetragen, bei der das in Abschnitt 6.4.2 vorgeschlagene Segmentierungsverfahren angewendet und den resultierenden Regionen ein zufälliger Salienzwert zugeordnet wurde. Wie zu sehen ist, erzielt der hier vorgeschlagene Ansatz das beste Ergebnis. Ein wichtiger Faktor ist hierbei die hierarchische Segmentierungsstrategie und die dadurch bedingte höhere Objektabdeckung. Wie jedoch durch die rein auf Segmentierung basierende Vergleichskurve deutlich wird, leistet auch die Messung der Salienz einen entscheidenden Beitrag. Auffallend ist, dass die Kurven der Vergleichsverfahren einen recht ähnlichen Verlauf haben. Dies ist dadurch

Kategorie	AP				
Person	59.3%	Flugzeug	55.7%	Bus	51.2%
Katze	42.6%	Motorrad	39.1%	Zug	36.7%
Bildschirm	35.7%	Fahrrad	33.7%	Hund	33.0%
Gemittelt	**30.0%**	Auto	26.3%	Schiff	25.8%
Pferd	25.3%	Sofa	23.3%	Schaf	21.9%
Vogel	21.5%	Stuhl	18.6%	Esstisch	18.5%
Flasche	12.8%	Kuh	12.0%	Topfpflanze	6.1%

Tabelle 7.3.2: Ergebnisse des Präsenzproblems auf dem VOC2012-Validierungsdatensatzes bei Auswertung der zehn salientesten Regionen.

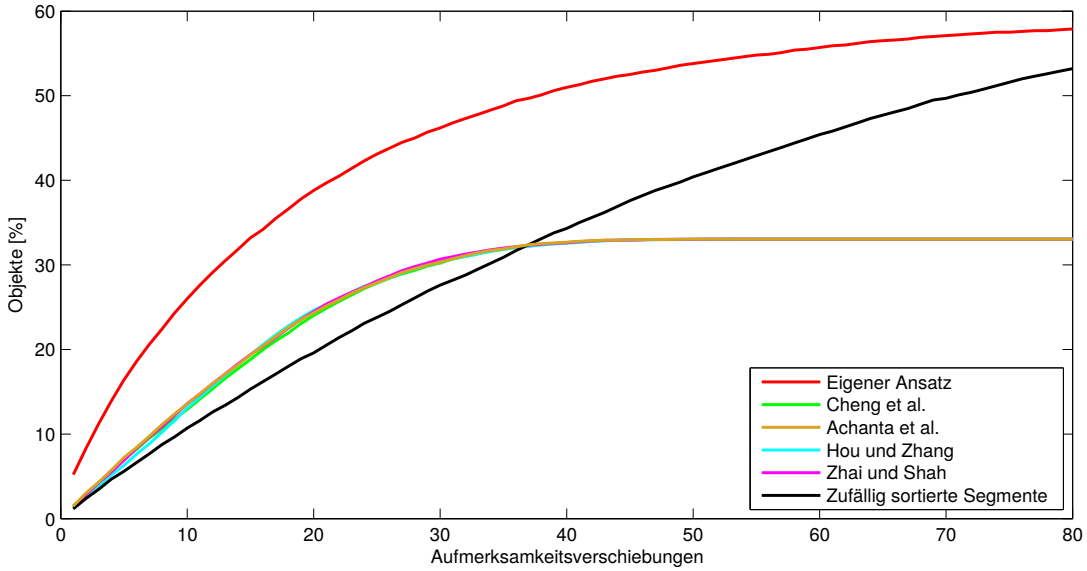

Abbildung 7.4.1: Vergleiche mit verschiedenen Bottom-Up-Verfahren auf dem VOC2012-Validierungsdatensatz.

begründet, dass die Verfahren bei einfacheren Szenen vergleichbar gute und bei Szenen mit komplexeren Konstellationen eher schlechte Ergebnisse erzielen.

Wie anhand der obigen Vergleiche deutlich wird, können die Ergebnisse vieler Bottom-Up-Verfahren der Vergangenheit auf schwierigen Datensätzen nicht überzeugen. In jüngerer Zeit haben deshalb Verfahren an Bedeutung gewonnen, die von der generischen Bottom-Up-Strategie abweichen und Modellwissen einsetzen, um bessere Ergebnisse zu erzielen. Entsprechende Top-Down-Strategien wurden in Abschnitt 5.1.4 betrachtet. Zwar entspricht der Top-Down-Ansatz nicht der in dieser Arbeit bevorzugten Problemauffassung, aufgrund der hohen Relevanz des Themas soll im Folgenden dennoch ein Vergleich gezogen werden. Hierzu wird das Verfahren von Alexe et al. [95] betrachtet, welches nach aktuellem Stand der Forschung ausgezeichnete Ergebnisse liefert. Das Verfahren verwendet ein Suchfensterstrategie und wurde in diesem Zusammenhang bereits in Abschnitt 5.1.3 betrachtet. Abbildung 7.4.2 zeigt den Vergleich auf dem VOC2007-Validierungsdatensatz. Wie zu erkennen ist, ist der Anteil an detektierten Objekten des Top-Down-Verfahrens innerhalb der ersten einhundert Aufmerksamkeitsverschiebungen zwischen 3.5% und 13% höher. Das Bottom-Up-Verfahren läuft danach aufgrund der begrenzten Anzahl an Objektkandidaten in die Sättigung, während das Top-Down-Verfahren aufgrund der Suchfensterstrategie stetig weiter ansteigt. Dies geschieht jedoch nur noch sehr langsam und bringt für den aufmerksamkeitsbasierten Ansatz, bei dem tendenziell nur wenige Inhalte ausgewertet werden sollen, im Grunde keinen Mehrwert mehr.

Das betrachtete Top-Down-Verfahren von Alexe et al. [95] nutzt verschiedene Merkmale bezüglich der Dichte von Kanten, dem Farbkontrast zur Umgebung,

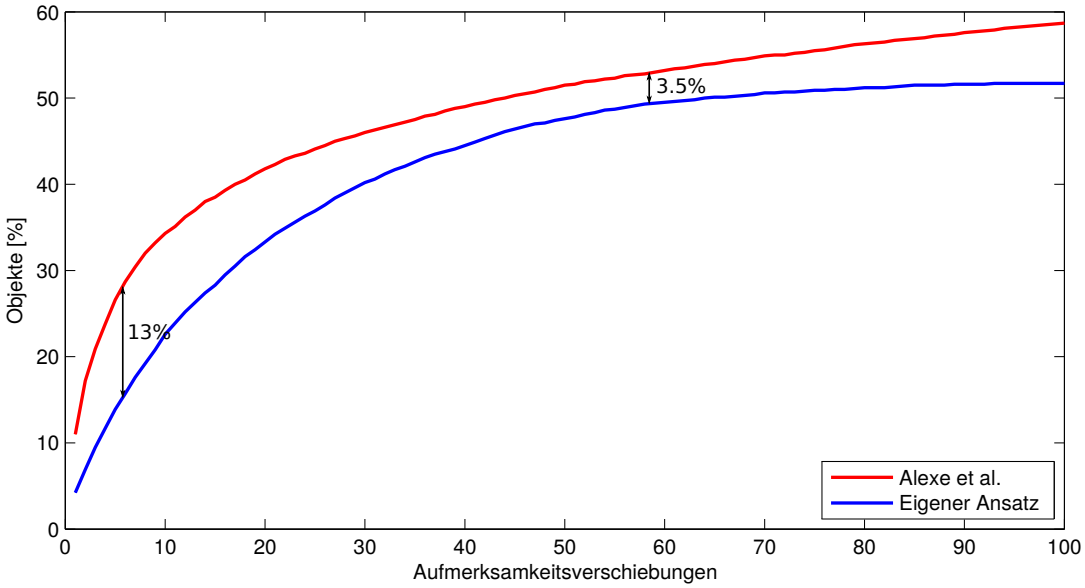

Abbildung 7.4.2: Vergleich zwischen dem hiesigen Bottom-Up-Verfahren (MV) und dem Top-Down-Verfahren von Alexe et al. [95] auf dem VOC2007-Validierungsdatensatz.

geschlossene Konturen, Position und Größe sowie den spektralen Rest nach Hou und Zhang [84]. Mittels einer Suchfensterstrategie wird eine große Anzahl an Objektkandidaten extrahiert und anhand dieser Merkmale bewertet. Dies geschieht durch die Verknüpfung mehrerer Bayes-Schätzer (Bishop [12], S. 152 ff.), bei denen Modellwissen eingesetzt wird, das aus Trainingsbildern gewonnen wird. Es werden im Sinne eines objektgenerischen Ansatzes ausschließlich Trainingsbeispiele problemfremder Objektkategorien verwendet. Das bedeutet, das Verfahren lernt nicht direkt die Merkmale von Zielobjekten, sondern das allgemeine Erkennen von Objektkonturen in verschiedenen Umgebungen. Dabei werden allgemeine Rahmenbedingungen ausgenutzt, wie sie bei typischen Fotoaufnahmen häufig vorliegen. Beispielsweise wird durch die statistische Auswertung von Größen und Positionen von Objekten im Bild ausgenutzt, dass bei Fotoaufnahmen die Kamera meistens annähernd parallel zum Boden ausgerichtet wird. Aus perspektivischen Gründen und aus Gründen der Schwerkraft gibt es dann bestimmte Positionen im Bild, an denen sich Objekte mit einer höheren Wahrscheinlichkeit aufhalten. Des Weiteren rückt der Fotograf häufig die interessanten Objekte der Szene bereit ins Zentrum der Aufnahme. Diese sind dann in der Regel fokussiert und weisen einen entsprechend hohen Kontrast auf, sodass sich hier weitere Rahmenbedingungen bezüglich der Konturen und Kantenstärken ergeben. Lässt man solche Annahmen in die Methodik des Detektors einfließen, kann man unter den entsprechenden Rahmenbedingungen zwar bessere Ergebnisse erzielen, schränkt die generische Einsetzbarkeit jedoch deutlich ein. Man könnte argumentieren, dass der Fotograf dem Detektor hier

bereits eine Teil der Arbeit abgenommen hat. Wird beispielsweise im Rahmen der Robotik ein Detektor eingesetzt, um interessanten Objekte zu finden, ist es kontraproduktiv anzunehmen, dass diese sich häufig bereits im Zentrum des Kamerabildes befinden und fokussiert sind.

Zusammenfassend lässt sich sagen, dass es Vorteile bringt, auf eine Top-Down-Strategie zu wechseln, wenn bestimmte Rahmenbedingungen bekannt sind. Jedoch reicht im vorliegenden Vergleich die Detektionsrate des Bottom-Up-Verfahrens, ohne den Vorteil der generischen Einsetzbarkeit einzuschränken, im relevanten Bereich auf bis zu 96.5% an die des Top-Down-Verfahrens heran.

7.5 DETAILBETRACHTUNGEN

Im Folgenden sollen einige Bildbeispiele zur Salienzdetektion betrachtet werden. Diese sind in den Abbildungen 7.5.4 bis 7.5.13 auf Seite 111 bis 120 dargestellt. Es handelt sich hierbei um die ersten zwanzig Bilder des VOC2011-Validierungsdatensatzes, die entsprechend mit 1 bis 20 nummeriert sind. In Reihe (a) ist jeweils das Originalbild und die einzelnen Stufen der Gauß-Pyramide zu sehen. Reihe (b) zeigt die zugehörigen Segmente. Reihe (c) zeigt die Ergebnisse für MV und Reihe (d) für GA. Die Darstellung der Segmente und Salienzwerte erfolgt hier analog zu Abbildung 7.3.1.

Bei der Bottom-Up-Variante, MV, wurden bei den Bildern 2, 4, 5, 7, 9, 10, 11, 12, 17, 18 und 20 die relevanten Inhalte im wesentlichen als salient eingestuft. Bei anderen Bildern kommt es hingegen zu Ablenkungen. So ist in Bild 1 der dargestellte Inhalt des Bildschirms und nicht der Bildschirm selbst salient. Bei Bild 3 wird schlichtweg der etwas beleuchtete Bereich im rechten Teil des Bildes als salient eingestuft. Bei den Bildern 6 und 13 ist es der Himmel und bei Bild 16 die Wasseroberfläche. Bild 14 weist viele uneinheitliche Elemente im Hintergrund auf, die von den Personen im Vordergrund ablenken. Interessanterweise sind dies alles Phänomene, die auch bei menschlichen Betrachtern zu Ablenkung führen können. Ein konzeptbedingtes Problem zeigt sich bei den Bildern 8 und 19. Sind zu wenig Umgebungsinformationen vorhanden, weil das Objekt den größten Teil der Bildfläche einnimmt, so wird in der Regel nicht das Objekt selbst als salient eingestuft, sondern lediglich Teile des Objekts, die im Vergleich zu dem Objekt als Ganzes hervorstechen. Bei Bild 8 ist dies der Reifen des Autos, bei Bild 19 die Pfote der Katze. Bei der Variante der gerichteten Aufmerksamkeit, GA, werden in der Regel Bereiche unterdrückt, die zwar hervorstechen, aber offensichtlich keine Objekte darstellen können. Bei den Bildern 6, 9 und 13 sind dies der Himmel, bei Bild 16 die Wasseroberfläche. Des Weiteren werden bei vielen Bildern die Objekte schneller detektiert als bei der MV-Variante. Bei den Beispielen wird dies vor allem am Pferd in Bild 4, am Flugzeug in Bild 6, am Auto in Bild 13 und an den beiden Vögeln in Bild 16 deutlich. Teilweise werden

$\varnothing\lvert\mathcal{R}\rvert$	$\overline{a_0}$	$a_0 > 0.5$
34.193 ± 5.496	0.429 ± 0.280	40.01%
66.610 ± 10.614	0.507 ± 0.257	52.94%
95.613 ± 14.662	0.538 ± 0.234	58.28%
121.904 ± 18.666	0.547 ± 0.223	59.43%

Tabelle 7.5.1: Durchschnittliche Anzahl an Segmenten und Segmentierungsgenauigkeit für MV auf dem VOC2012-Validierungsdatensatz bei unterschiedlichen Konfigurationen der dynamischen Schwellwertanpassung.

auch Objekte aus problemfremden Kategorien detektiert. Beispielsweise ist dies in Bild 12 bei der Wanduhr der Fall.

Bezüglich der Detektionsgenauigkeit ergeben sich gelegentlich Probleme, wie sie für Bottom-Up-Segmentierungsverfahren nicht unüblichen sind. So können Teile des Hintergrunds zum Objekt gerechnet (Bild 9) oder Teile des Objekts abgeschnitten werden (Bild 20). Des Weiteren ist es möglich, dass ähnliche benachbarte oder perspektivisch überlappende Objekte zu einer Region zusammengefasst werden (Bild 12), oder dass ein Objekt in mehrere Teile fragmentiert wird. Letzteres wurde bereits in Abschnitt 7.3 im Zusammenhang mit der Objektkategorie *Person* thematisiert. Ist die trennende Kontur innerhalb des Objekts stärker ausgeprägt als die äußere Kontur des Objekts, löst auch der hierarchische Segmentierungsansatz das Problem nicht. Bei den Bildern 18 und 20 lässt sich erkennen, dass zwar die einzelnen Teile der Person als salient eingestuft werden, das Objekt selbst wurde aufgrund der Fragmentierung jedoch nicht detektiert. In Kapitel 8 werden ausblickend noch einige Anregungen für zukünftige Verbesserungen und Weiterentwicklungen gegeben, die einige der hier dargestellten Probleme adressieren.

Im Folgenden soll der Blick noch auf einige weitere Details gerichtet werden. Als Erstes wird die durchschnittliche Anzahl der Objektkandidaten pro Bild betrachtet. Bei den bisher präsentierten Ergebnissen lag die Anzahl bei 95.613 ± 14.662. Ein interessanter Punkt ist der Einfluss auf die Ergebnisse, wenn man diese Zahl durch einen engeren Rahmen bei der dynamischen Schwellwertanpassung herauf- oder herunterregelt. In Tabelle 7.5.1 werden Segmentierungsergebnisse für verschiedene Konfigurationen diesbezüglich angegeben. Wie zu sehen ist, steigt die Segmentierungsgenauigkeit und die Abdeckung der Objekte mit steigender Zahl an Regionen. Die Zahl an Regionen nimmt dabei jedoch aufgrund von zunehmender Übersegmentierung überproportional zu. Dies ist problematisch, da hierdurch die durchschnittliche Qualität der Objektkandidaten abnimmt. Es wird nämlich bei einer steigenden Anzahl an Segmenten tendenziell schwieriger, allein durch das Salienzkriterium gute Objekthypothesen aufzustellen. Dieser

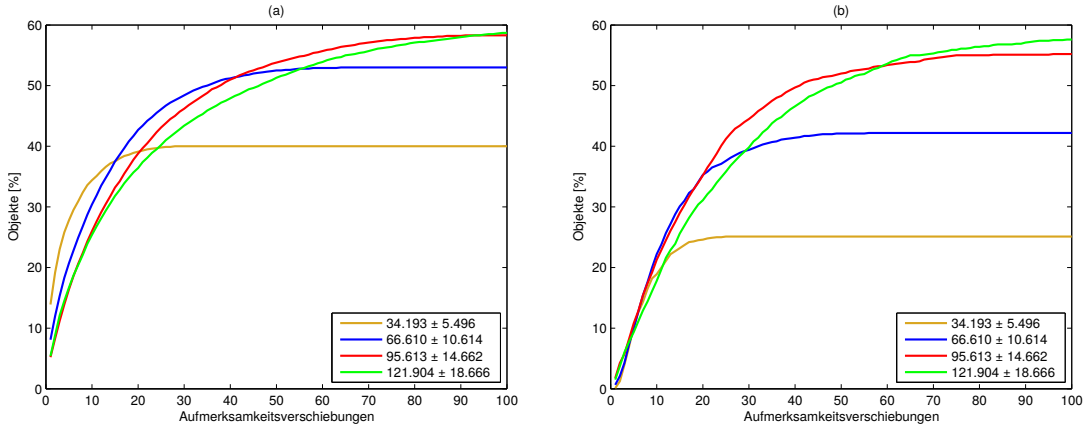

Abbildung 7.5.1: Detektionsergebnisse auf dem VOC2012-Validierungsdatensatz für MV bei unterschiedlichen Segmentierungsschwellwerten. (a) Alle Objekte. (b) Nur kleine Objekte.

Zusammenhang wird in Abbildung 7.5.1 (a) verdeutlicht. Der Vergleich der roten und grünen Kurve zeigt, dass trotz einer höheren Objektabdeckung schlechtere Detektionsergebnisse erzielt werden. Umgekehrt lässt sich die durchschnittliche Qualität der Objektvorhersagen durch eine Reduktion der Segmentanzahl auch nicht beliebig steigern. Zwar wird der Kurvenanstieg bei abnehmender Segmentanzahl steiler, jedoch werden dann aufgrund einer zunehmenden Untersegmentierung kleinere Objekte häufiger übersehen. Dies wird in Abbildung 7.5.1 (b) verdeutlicht, bei der nur die Ergebnisse für kleine Objekte dargestellt sind. Als klein werden hier Objekte eingestuft, deren Bounding-Boxen weniger als 10% der Bildfläche einnehmen. Wie zu erkennen ist, werden die Detektionsergebnisse bei weniger Segmenten für kleine Objekte schlechter.

Ein weiterer interessanter Punkt ist der Einfluss der Merkmalsintegration auf die Detektionsergebnisse. Wie in Abschnitt 4.1 erläutert wurde, handelt es sich bei der Merkmalsintegration um einen anhand von Verhaltensexperimenten gut belegbaren Bestandteil der menschlichen Aufmerksamkeit. Das Konzept wurde, wie in Abschnitt 6.4.3 beschrieben wurde, für das hier vorgeschlagene Verfahren adaptiert. Um den Einfluss dieser Strategie auf das Gesamtergebnis zu evaluieren, werden in Abbildung 7.5.2 die Detektionsergebnisse bei Betrachtung aller Salienzmerkmale mit Ergebnissen verglichen, bei denen einzelne Merkmalstypen ausgelassen wurden. Wie in Teil (a) zu sehen ist, werden bis zu 1.3% weniger Objekte detektiert, wenn die Farbmerkmale bei der Merkmalsintegration nicht berücksichtigt werden. Teil (b) zeigt analog den Vergleich bei Auslassung der Orientierungsmerkmale. Hier werden bis zu 0.7% weniger Objekte detektiert. Bei einzelnen Objektkategorien fällt der Unterschied größer aus. Teil (c) zeigt den Vergleich bei Auslassung der Farbmerkmale für die Objektkategorie *Stuhl*. Hier Beträgt der Abstand der Kurven bis zu 2.7%. Wie in Teil (d) zu sehen

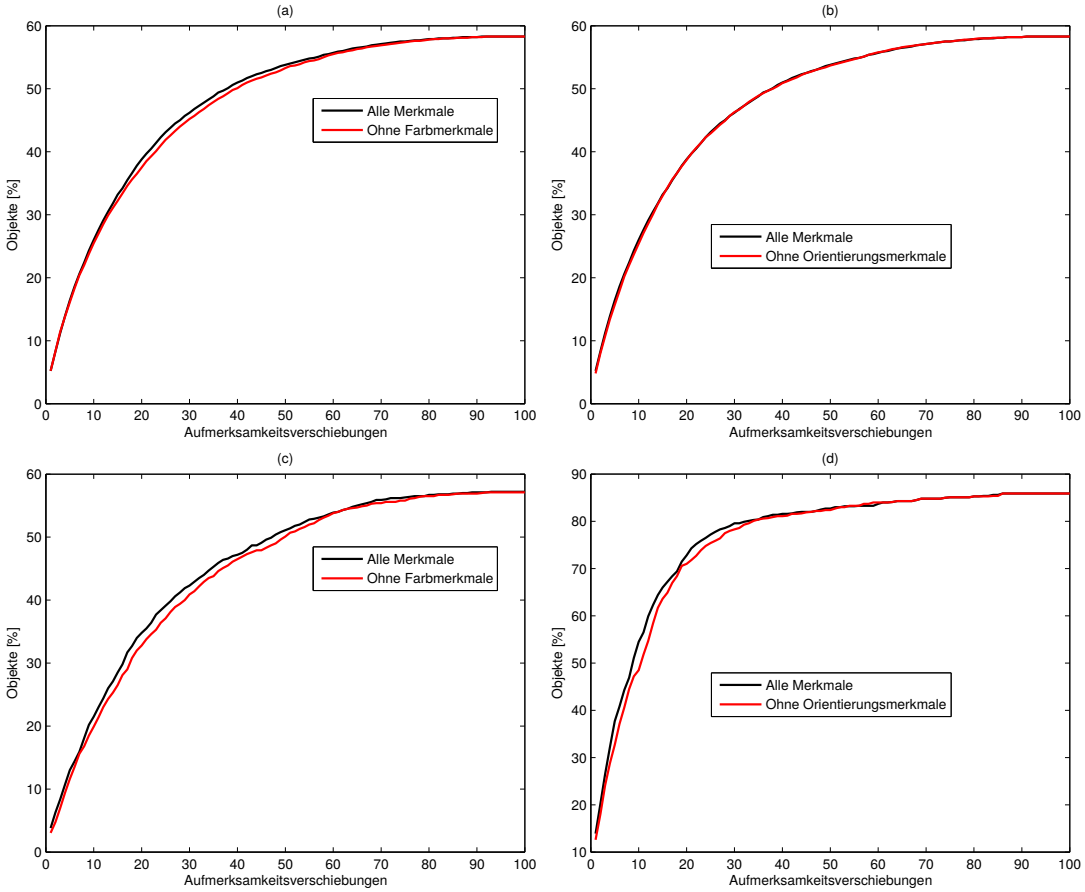

Abbildung 7.5.2: Einfluss der Merkmalsintegration: (a) Vergleich ohne Farbmerkmale, alle Objektkategorien. (b) Vergleich ohne Orientierungsmerkmale, alle Objektkategorien. (c) Vergleich ohne Farbmerkmale, Objektkategorie *Stuhl*. (d) Vergleich ohne Orientierungsmerkmale, Objektkategorie *Katze*. Dargestellt sind jeweils die Ergebnisse für MV auf dem VOC2012-Validierungsdatensatz.

ist, beträgt der Abstand für die Objektkategorie *Katze* bei Auslassung der Orientierungsmerkmale sogar 6%.

Abschließend sollen noch die Ergebnisse hinsichtlich unterschiedlicher Objekteigenschaften betrachtet werden. Hierbei werden die zusätzlichen Angabe der VOC2012-Datenbank zur Perspektive und Sichtbarkeit der annotierten Objekte ausgewertet (siehe Abschnitt 7.1). Des Weiteren werden auch die Projektionsgrößen der Objekte betrachtet. Dies geschieht, indem die Größe der Bounding-Box des jeweiligen Objekts ins Verhältnis zur Bildgröße gesetzt wird. Anhand dieses Verhältnisses werden die Objekte hier in die Kategorien *sehr klein*, *klein*, *mittel*, *groß* und *sehr groß* eingeteilt. Als Grenzwerte für diese Einteilung wurden 5%, 10%, 25% und 50% der Bildgröße festgelegt. Tabelle 7.5.2 zeigt zunächst die

	Objekte	$\overline{a_0}$	$a_0 > 0.5$
Größe			
Sehr klein ($< 5\%$)	6067	0.386 ± 0.235	31.10%
Klein ($5\% - 10\%$)	1994	0.546 ± 0.177	55.06%
Mittel($10\% - 25\%$)	2946	0.590 ± 0.163	67.58%
Groß ($25\% - 50\%$)	2565	0.622 ± 0.142	78.44%
Sehr groß ($> 50\%$)	2215	0.783 ± 0.131	100%
Perspektive			
Frontal	3759	0.572 ± 0.216	63.95%
Rückseitig	1192	0.473 ± 0.244	47.40%
Links	2202	0.537 ± 0.219	58.13%
Rechts	2246	0.529 ± 0.225	57.17%
Sichtbarkeit			
Am Rand abgeschnitten	8288	0.567 ± 0.237	62.38%
Teilweise verdeckt	6651	0.522 ± 0.215	53.84%
Schwer zu erkennen	1946	0.383 ± 0.279	33.92%
Gesamt	**15787**	$\mathbf{0.538 \pm 0.234}$	**58.28%**

Tabelle 7.5.2: Segmentierungsergebnisse für MV auf dem VOC2012-Validierungsdatensatz bei Betrachtung unterschiedlicher Objekteigenschaften.

Ergebnisse der Segmentierung. Bezüglich der Objektgröße lässt sich feststellen, dass die Ergebnisse für große Objekte besser ausfallen, als für kleine Objekte. Der Grund besteht darin, dass für kleine Objekte der Lösungsraum größer ist. Sie werden in der Regel auf einer feinen Skalierungsstufe adressiert. Bei vielen kleineren Regionen gibt es jedoch bedeutend mehr Möglichkeiten, das Bild aufzuteilen, als bei wenigen großen Regionen. Nimmt ein Objekt mehr als die Hälfte der Bildfläche ein, wird es aufgrund der hierarchischen Segmentierungsstrategie immer eine Region geben, die das Kriterium $a_0 > 0.5$ erfüllt. Bezüglich der Objektperspektive lässt sich feststellen, dass die Trefferquote für Objekte aus der Frontalansicht um 16% besser ist, als für Objekte aus der Rückansicht. Dies kann zum Teil an den visuellen Eigenschaften der Objekte liegen. Es muss aber auch berücksichtigt werden, dass Objekte aus der Rückansicht tendenziell eher zufällig in den Hintergrund des Bildes geraten, da sie nicht das eigentliche Ziel der Aufnahme darstellen. Für Objekte, die am Bildrand abgeschnitten werden, ist die Segmentierungsgenauigkeit besser als im Durchschnitt. Der Grund hierfür ist, dass die annotierte Bounding-Box eines abgeschnittenen Objekts mit der Bounding-Box des korrespondierenden Segments am entsprechenden Bildrand meist genau übereinstimmt. Im Durchschnitt schlechter hingegen sind die Er-

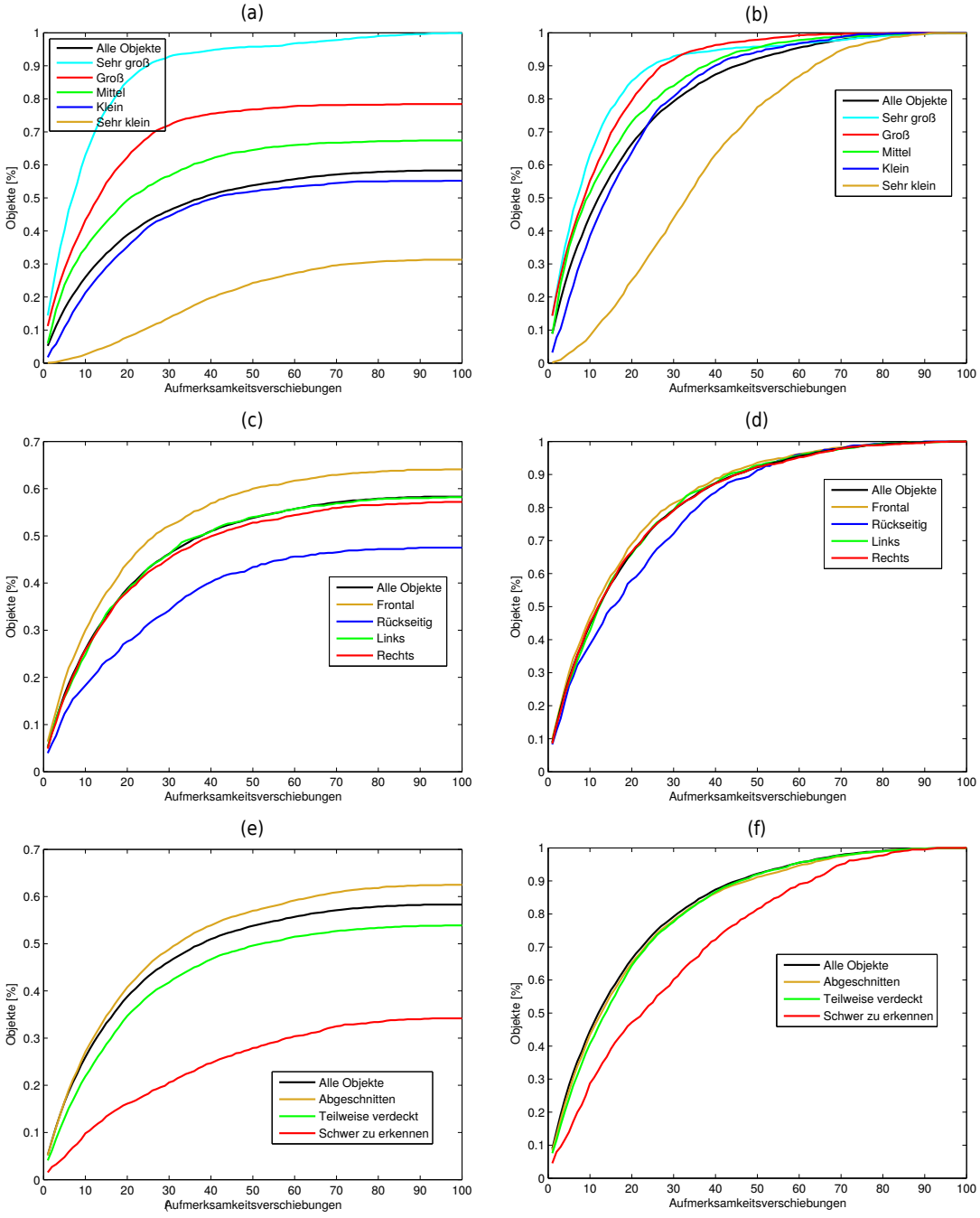

Abbildung 7.5.3: Detektionsergebnisse für MV auf dem VOC2012-Validierungsdatensatz bei Betrachtung unterschiedlicher Objekteigenschaften: (a) Größe. (b) Größe, nur passende Segmente. (c) Perspektive. (d) Perspektive, nur passende Segmente. (e) Sichtbarkeit. (f) Sichtbarkeit, nur passende Segmente.

gebnisse für teilweise verdeckte Objekte. Dies liegt vor allem daran, dass solche Objekte teilweise durch die verdeckenden Elemente in mehrere Bereiche zerlegt

werden und dann bei der Segmentierung entsprechend in mehrere Segmente unterteilt werden. Des Weiteren sind auch die Ergebnisse für Objekte, die von den Bearbeitern als besonders schwer zu erkennen eingestuft worden sind, wesentlich schlechter als im Durchschnitt.

In Abbildung 7.5.3 werden die auf den Segmentierungsergebnissen aufbauenden Detektionsergebnisse dargestellt. Um den Einfluss der Segmentierung zu verdeutlichen, sind jeweils auf der rechten Seite zusätzlich die Ergebnisse dargestellt, bei denen nur die Objekte mit passenden Segmenten berücksichtigt wurden. Bei (b) lässt sich hieraus erkennen, dass das Salienzkriterium bei Objekten, die eine kritische Größe unterschreiten, schlechter funktioniert, als bei größeren Objekten. In geringerem Umfang gilt dies auch für Objekte aus rückwärtiger Perspektive (d) und für Objekte, die als schwer zu erkennen eingestuft wurden (f).

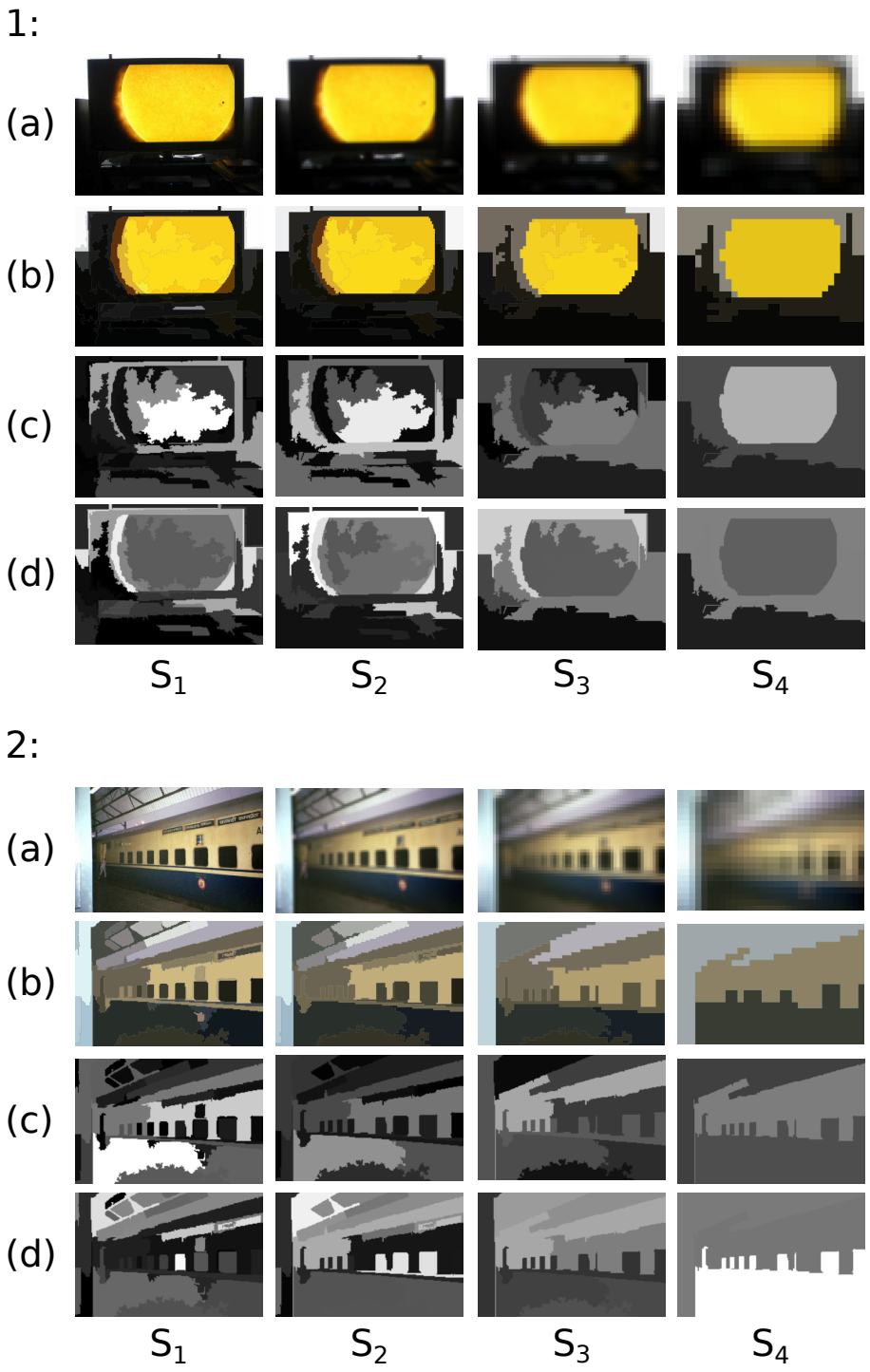

Abbildung 7.5.4: Ergebnisse der Salienzdetektion für das 1. und 2. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

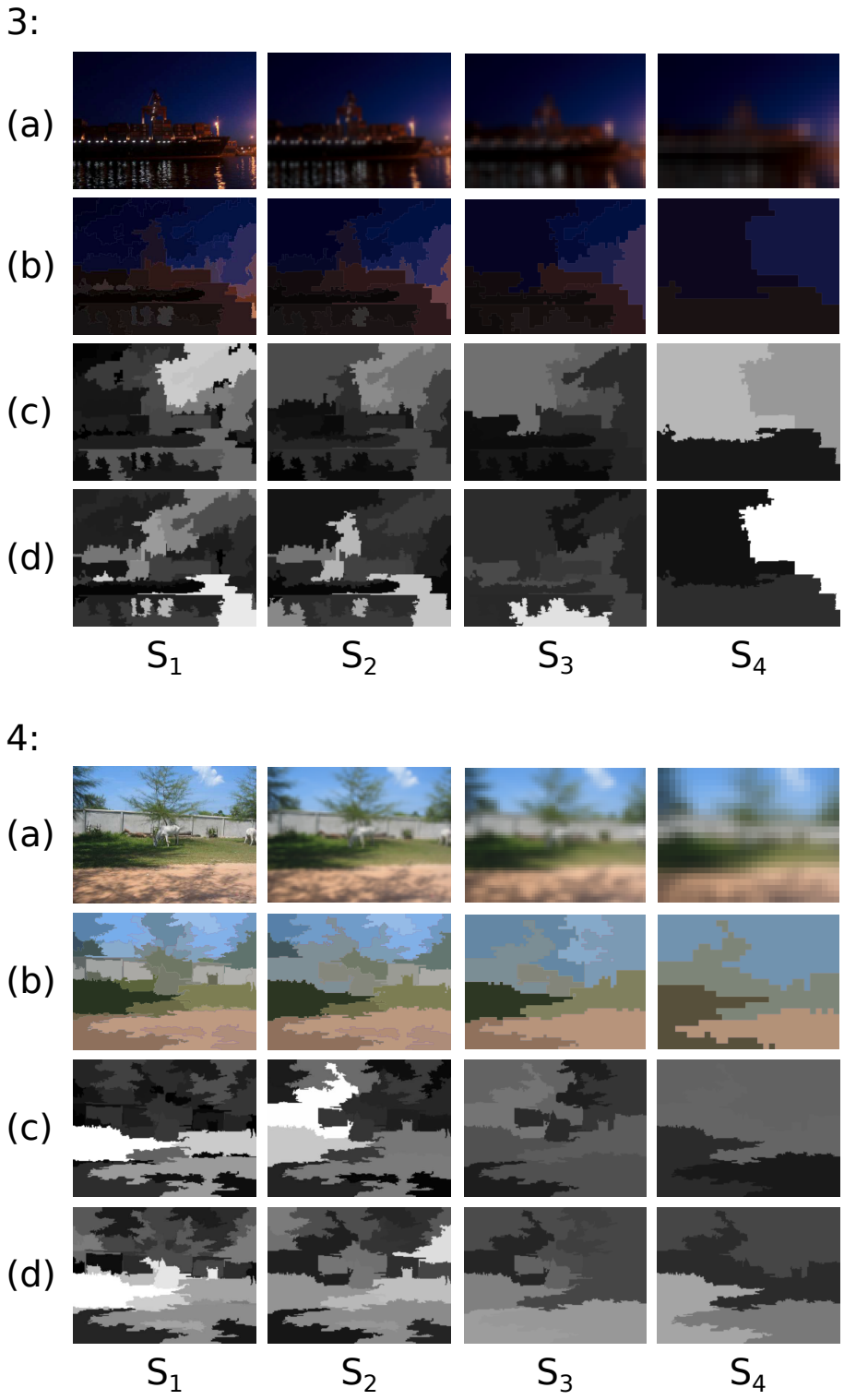

Abbildung 7.5.5: Ergebnisse der Salienzdetektion für das 3. und 4. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

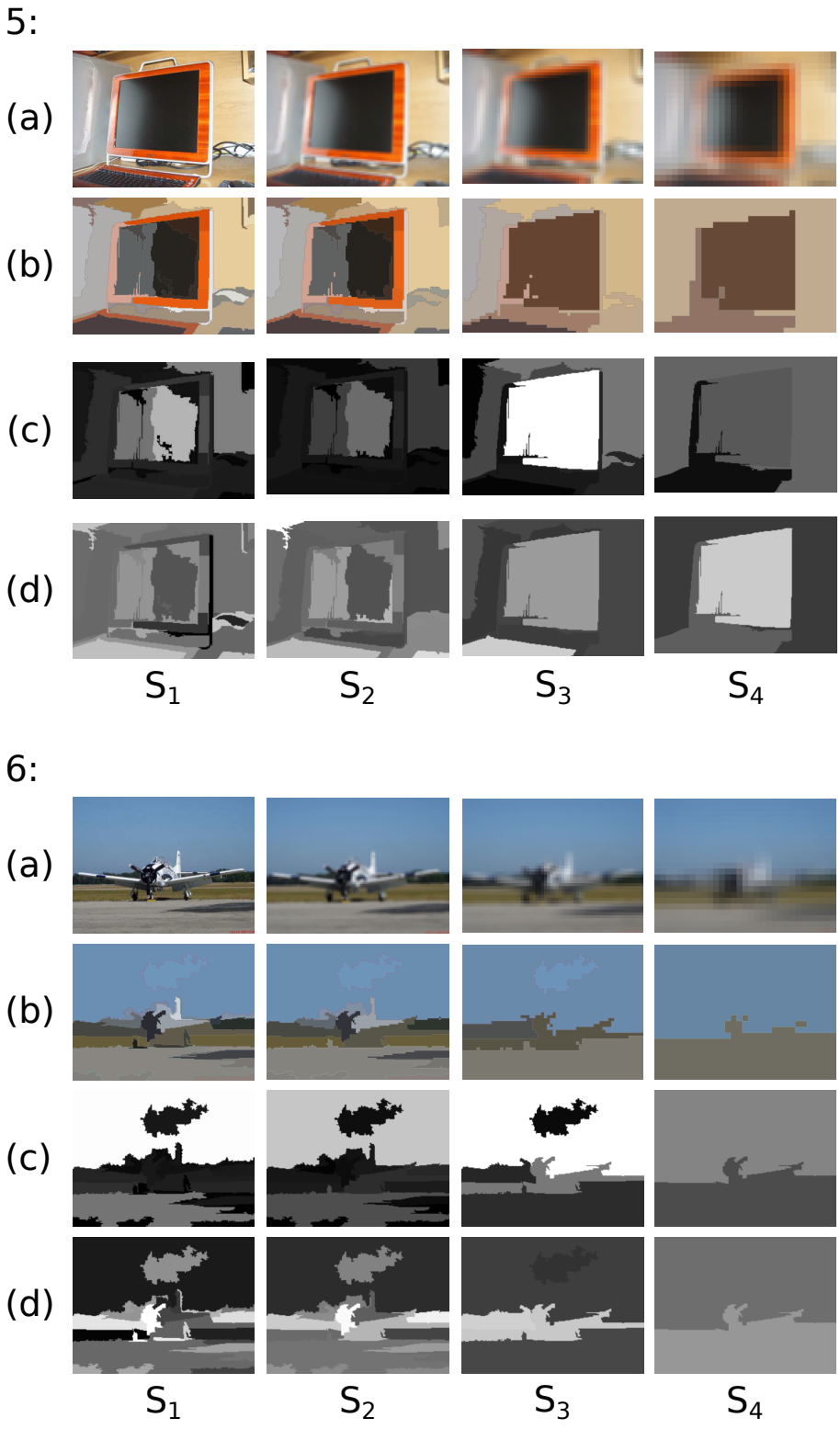

Abbildung 7.5.6: Ergebnisse der Salienzdetektion für das 5. und 6. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

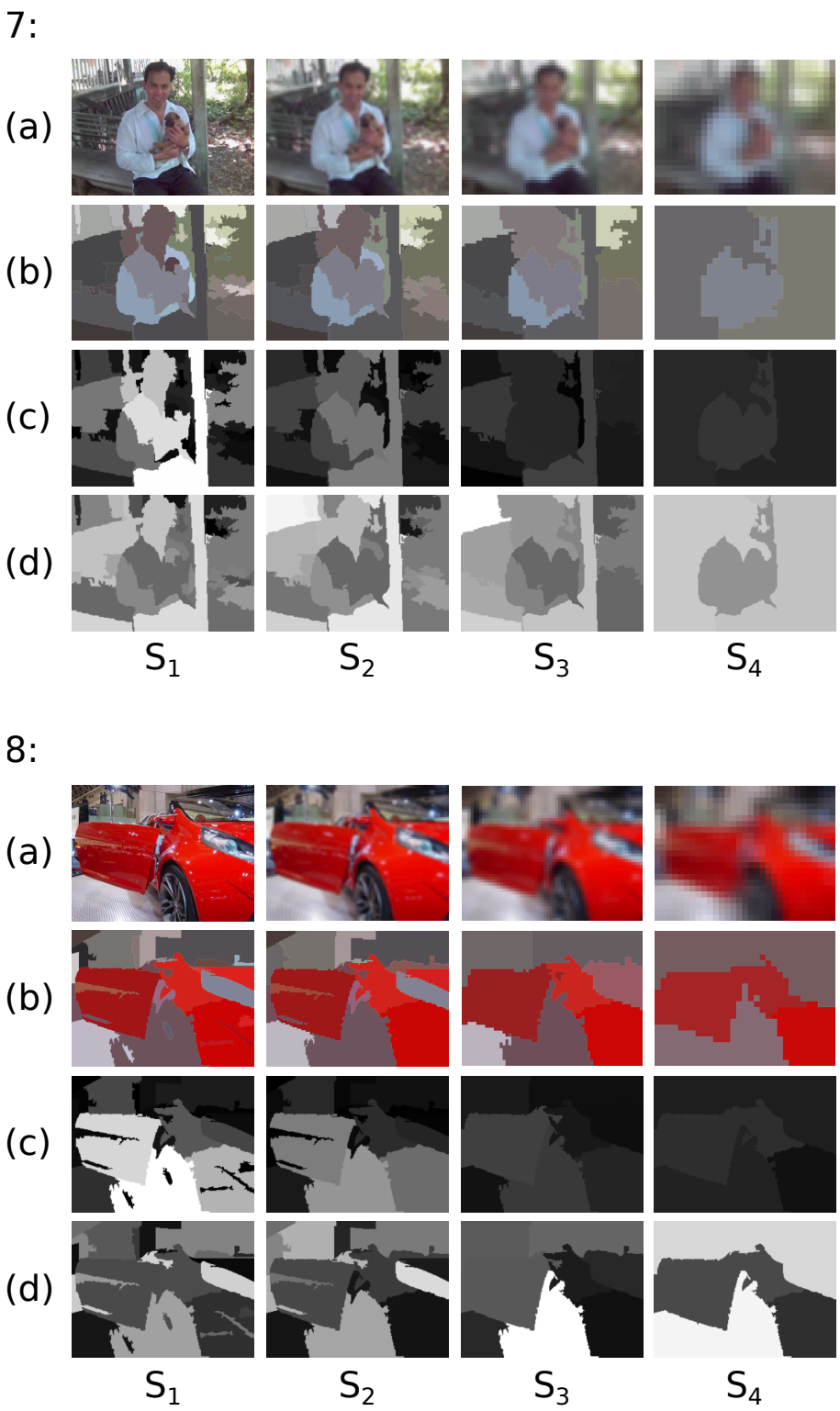

Abbildung 7.5.7: Ergebnisse der Salienzdetektion für das 7. und 8. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

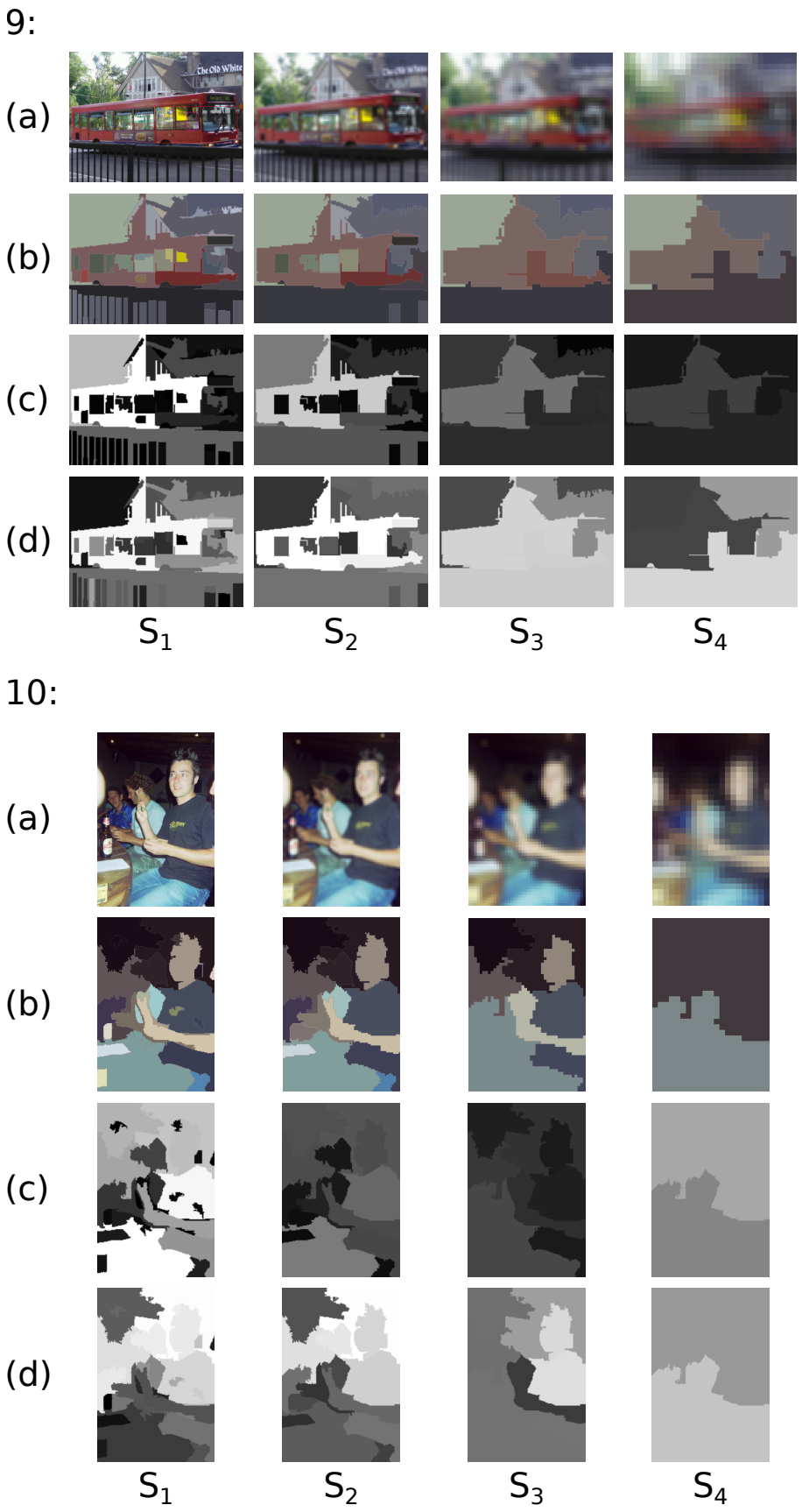

Abbildung 7.5.8: Ergebnisse der Salienzdetektion für das 9. und 10. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

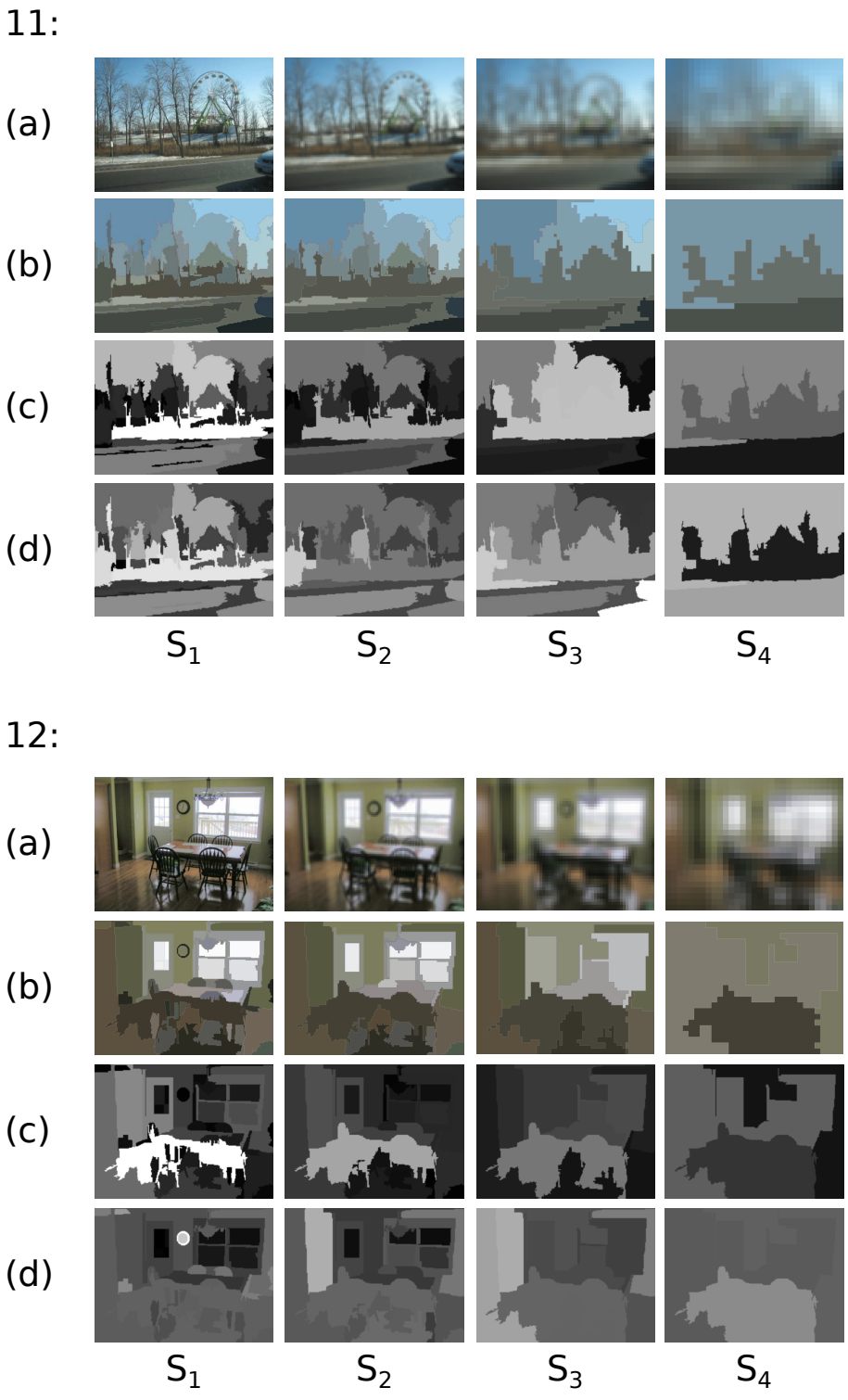

Abbildung 7.5.9: Ergebnisse der Salienzdetektion für das 11. und 12. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

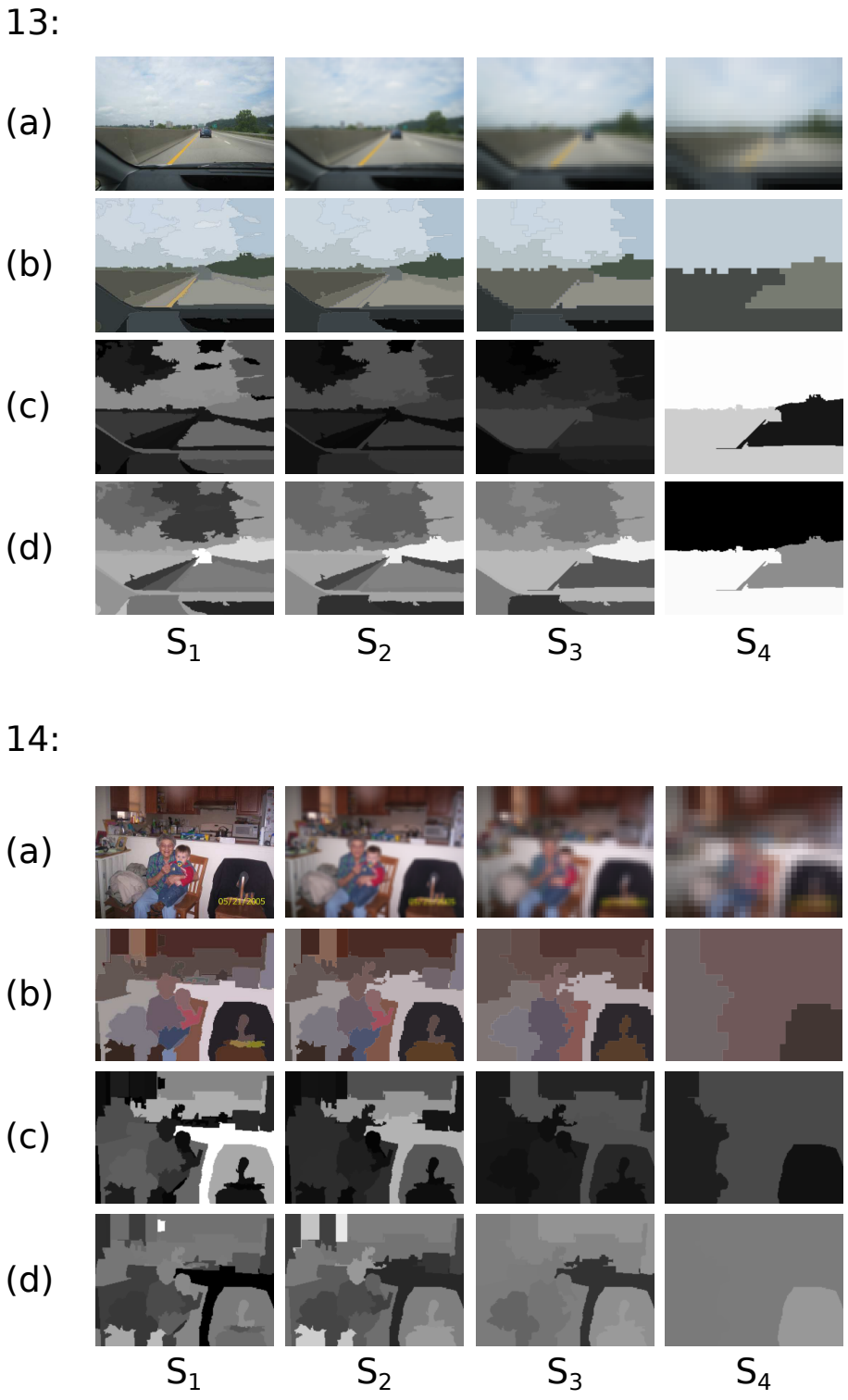

Abbildung 7.5.10: Ergebnisse der Salienzdetektion für das 13. und 14. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

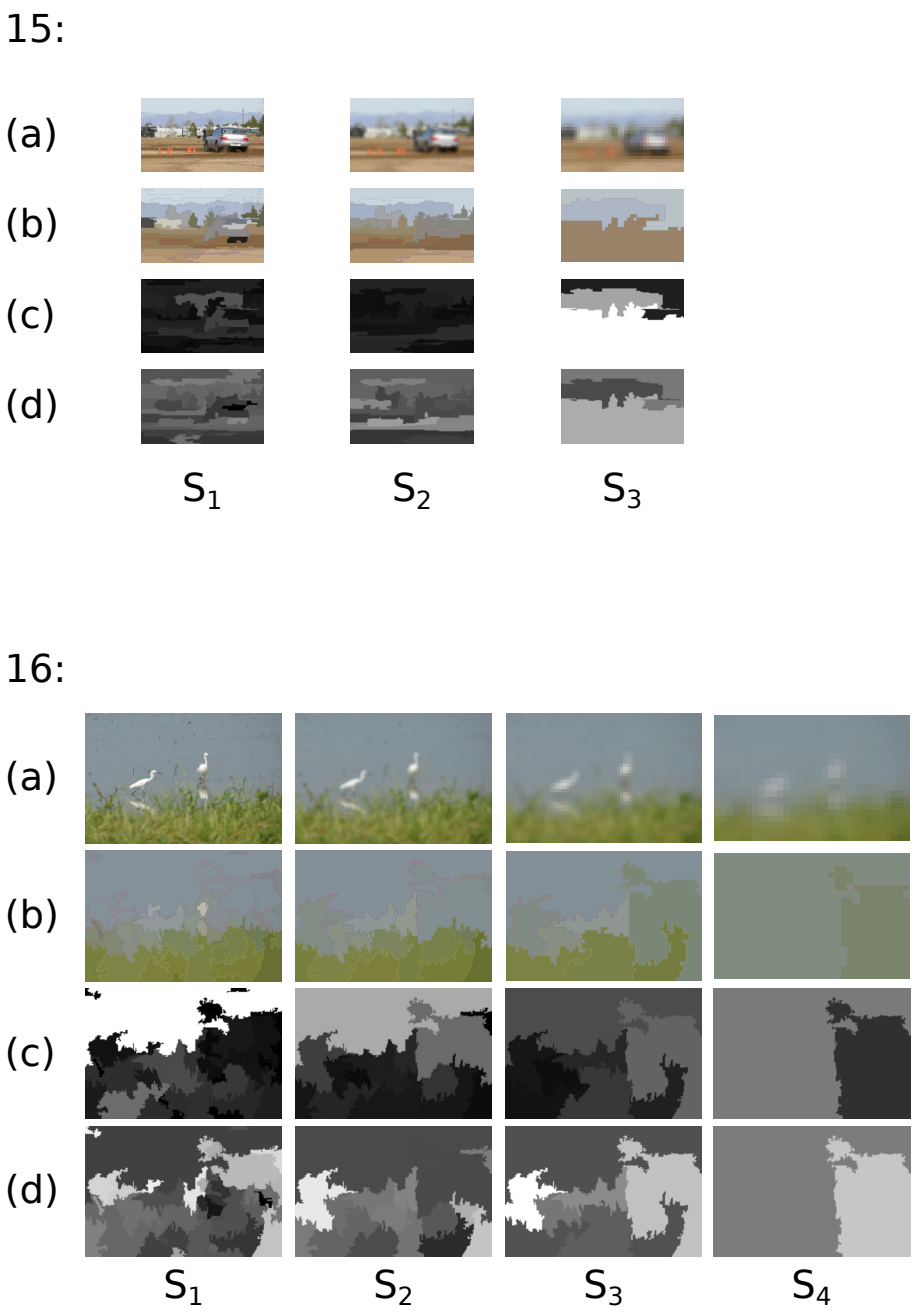

Abbildung 7.5.11: Ergebnisse der Salienzdetektion für das 15. und 16. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

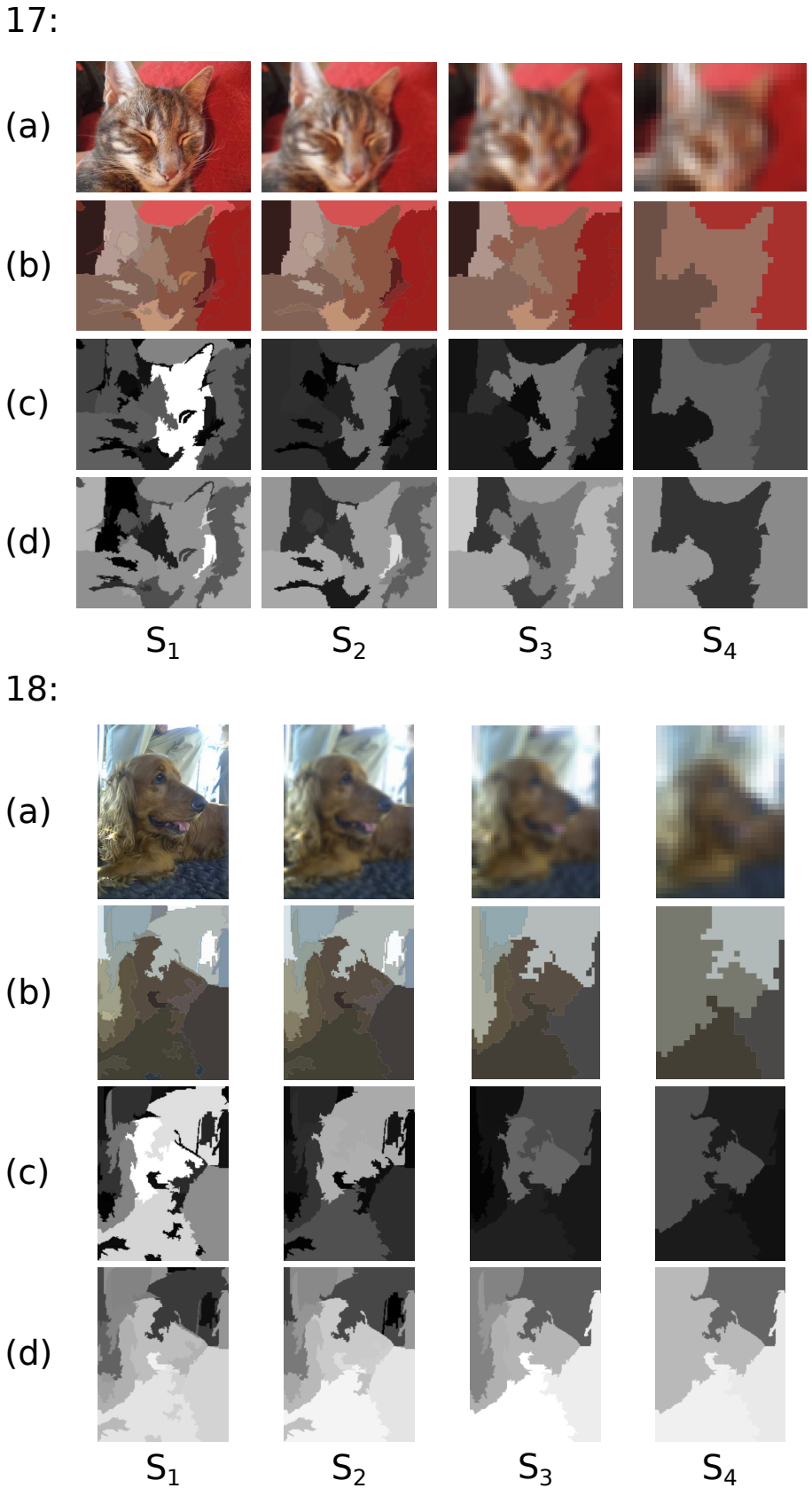

Abbildung 7.5.12: Ergebnisse der Salienzdetektion für das 17. und 18. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

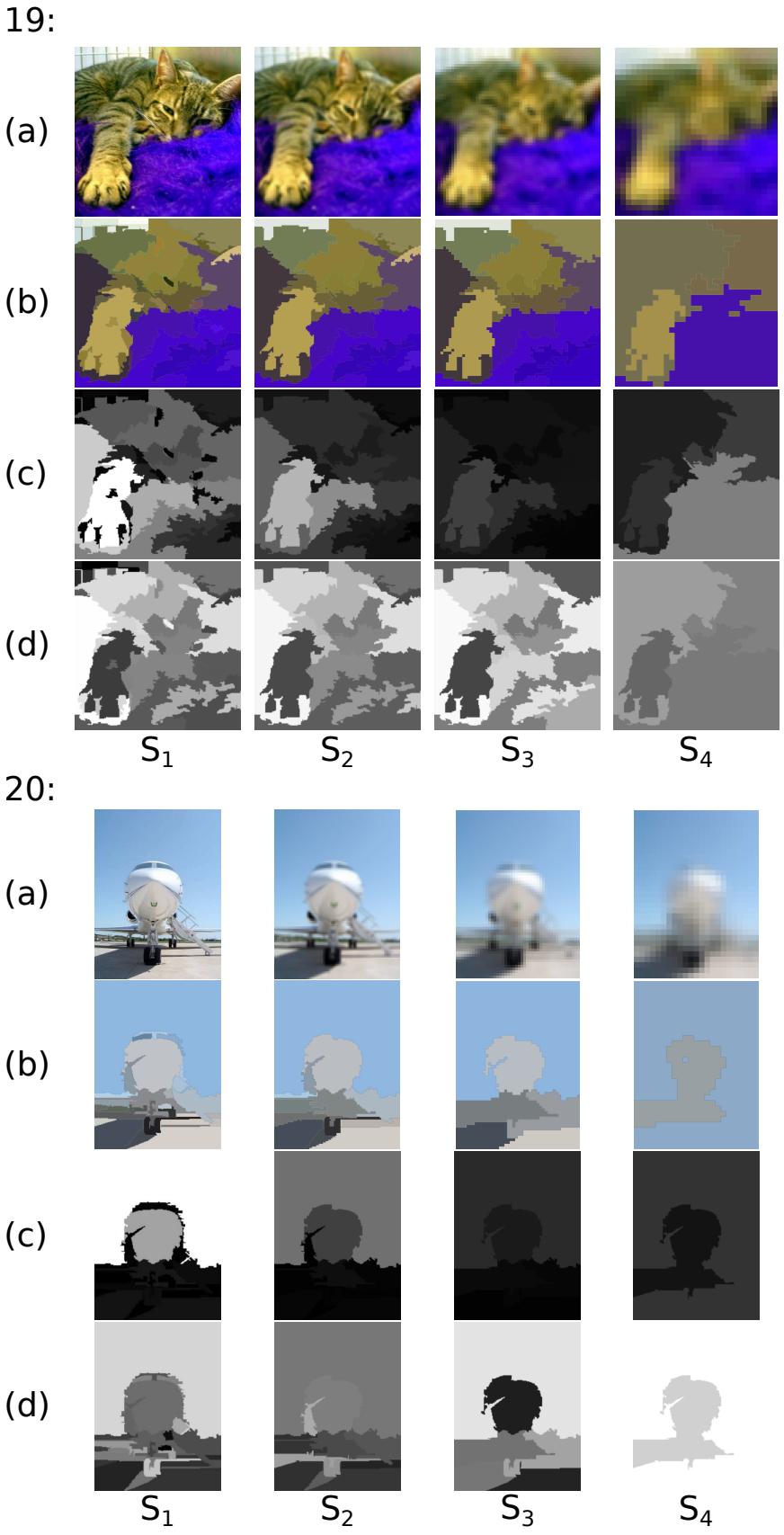

Abbildung 7.5.13: Ergebnisse der Salienzdetektion für das 19. und 20. Bild des VOC2012-Validierungsdatensatzes. (a) Gauß-Pyramide. (b) Segmente. (c) MV. (d) GA.

8

SCHLUSSWORT

Diese Arbeit hatte die aufmerksamkeitsbasierte Objektdetektion zum Gegenstand. Motiviert wurde das Thema als eine Alternative zur klassischen Objektdetektion. Durch einen aufmerksamkeitsbasierten Ansatz sollte eine schnelle und aufwandsarme Auswertung der interessanten Szeneninhalte ermöglicht werden. Hierzu wurde die Objektdetektion als wichtiges Anwendungsbeispiel betrachtet, was sich auf der These gründete, dass die interessanten Objekte einer Szene visuell salient sind. Klassische Detektionsverfahren sind im Allgemeinen rechenintensiv und skalieren schlecht mit der Anzahl der betrachteten Objektkategorien. Die zentrale Fragestellung war es deshalb, ob sich Salienz als Kriterium zur Lokalisierung von Objekten in komplexen Szenen nutzen lässt. Zu diesem Zweck wurden zunächst Modelle und Theorien zur menschlichen Aufmerksamkeit betrachtet. Anschließend wurden mathematische Aufmerksamkeitsmodelle aus der Literatur betrachtet. Aufbauend darauf wurde ein eigenes Aufmerksamkeitsmodell vorgeschlagen, welches geschlossene Konturen detektiert und deren Salienz bewertet. Das Verfahren wurde auf einem schwierigen Benchmark evaluiert, das viele Bilder mit komplexen Inhalten aufweist. Es konnte gezeigt werden, dass das vorgeschlagene Bottom-Up-Verfahren auf dem betrachteten Benchmark bekannten Bottom-Up-Ansätzen aus der Literatur überlegen ist. Wichtige Faktoren, die hierzu beitragen, sind die hierarchische Segmentierungsstrategie und die Verwendung geeigneter regionaler Salienzmerkmale. Des Weiteren wurde ein Vergleich mit einem objektgenerischen Top-Down-Verfahren gezogen, welches nach aktuellem Stand der Wissenschaft hervorragende Ergebnisse erzielt. Es wurde gezeigt, dass in dem für den aufmerksamkeitsbasierten Ansatz relevanten Bereich die Detektionsrate des vorgeschlagenen Bottom-Up-Verfahrens auf bis zu 96.5% an die des Vergleichsverfahrens heranreicht. Nicht ausgeschlossen ist, dass dieser Abstand im Rahmen zukünftiger Forschung noch weiter reduziert werden kann.

Im Folgenden sollen ausblickend noch einige Anregungen für weiterführende Forschungsthemen gegeben werden. Eine naheliegende Verbesserungsmöglichkeit ist es, bei der Merkmalsintegration weitere Merkmalstypen zu betrachten. Hierbei kommen vor allem Merkmale auf Basis von Bewegung und Tiefe in Frage. Entsprechende Merkmalskarten ließen sich bei der Betrachtung von Videosequenzen beziehungsweise Stereobildern ermitteln. Bewegungs- und Tiefeninformationen können auch eingesetzt werden, um die Qualität der Segmentierung zu verbessern.

Ein weiteres Thema für zukünftige Forschung kann es sein, das Zusammenspiel zwischen der präattentiven und der attentiven Phase der menschlichen Wahrneh-

mung stärker zu beleuchten. Hierbei sind beispielsweise Ansätze interessant, bei denen die präattentive Phase durch ein Feedback der attentiven Phase beeinflusst wird. Auf diese Weise kann sich ein Wechsel zwischen rein datengetriebener und gerichteter Aufmerksamkeit vollziehen. In dieser Arbeit wurde bereits ein Verfahren zur gerichteter Aufmerksamkeit vorgeschlagen, welches zu diesem Zweck eingesetzt werden kann. Können im ersten Schritt durch die Auswertung weniger salienter Inhalte bereits bestimmte Rückschlüsse auf die Szene gezogen werden, kann dynamisch auf ein gerichtetes Aufmerksamkeitsverfahren gewechselt werden, welches über entsprechendes Modellwissen verfügt. Hierbei wäre es auch möglich, Methoden der Szenenerkennung einzusetzen. Beispielsweise erwartet man in einer Indoor-Szene andere Objekte anzutreffen als in einer Outdoor-Szene. Stellen sich erst während der Erkundung einer Szenen bestimmte Inhalte als relevant heraus, wäre der Einsatz eines Trainingsverfahrens zur Laufzeit denkbar. Ein entsprechendes Aufmerksamkeitsmodell würde sich dynamisch an seine Umgebung anpassen und den Wechsel zwischen datengetriebener und gerichteter Aufmerksamkeit schrittweise vollziehen.

Weitere Verbesserungsmöglichkeiten können sich auf die attentive Phase der Wahrnehmung beziehen. In dieser Arbeit wurde ein Bag-of-Features-Ansatz favorisiert, da sich dieser generisch auf verschiedenste Objektklassen anwenden lässt. Es kann aber auch sinnvoll sein, teilebasierte Methoden zur Auswertung salienter Regionen einzusetzen. Hierbei könnten nachträglich Probleme adressiert werden, wie sie bei der vorangehenden Bottom-Up-Detektion typischerweise auftreten. Beispielsweise könnte eine Region daraufhin untersucht werden, ob sie nur einen Teil eines größeren Objekts zeigt, oder ob sie mehrere Objekte beinhaltet. Ein solches Vorgehen kann sich auf solche Objekte konzentrieren, die in der Praxis besonders häufig anzutreffen sind. Hierzu dürften insbesondere Personen zählen. Wird beispielsweise im ersten Schritt ein Gesicht detektiert, können im nächsten Schritt adjazente Regionen ausgewertet werden, um die Person als Ganzes zu detektieren. Eine weitere Verbesserungsmöglichkeit besteht darin, der Bottom-Up-Segmentierung der präattentiven Phase ein komplexeres, modellbasiertes Segmentierungsverfahren im attentiven Pfad zur Verfeinerung der salientesten Regionen nachzuschalten. Hierdurch könnte die Detektionsgenauigkeit und somit auch die Detektionsrate verbessert werden.

Die hier aufgezeigten Erweiterungs- und Verbesserungsmöglichkeiten sollen zur weiteren Forschung motivieren. Die visuelle Aufmerksamkeit im Allgemeinen und die aufmerksamkeitsbasierte Objektdetektion im Speziellen sind spannende und lohnenswerte Themengebiete, die vielfältige Ansatzpunkte für weitere Forschungsarbeiten bieten.

Die in dieser Arbeit vorgestellten Methoden sind das Ergebnis stetiger Weiterentwicklungen. Zwischenstände, die sich inhaltlich mit den hier präsentierten Methoden überschneiden, sind bereits in [123] sowie [124] veröffentlicht worden. Da diese Publikationen unter der Beteiligung von Koautoren entstanden sind, sollen im Folgenden die Beiträge des Autors dieser Arbeit von den Beiträgen der anderen Autoren abgegrenzt werden.

In [123] wird ein Verfahren vorgeschlagen, welches Modelle zur Objektdetektion aus unzuverlässigen Quellen erlernt. Der Beitrag des hiesigen Autors an dieser Veröffentlichung besteht in der Methodik zur Bottom-Up-Detektion und anschließenden Klassifizierung von Objektvorschlägen. Es handelt sich dabei um das zweistufiges Modell aus einer Bottom-Up-Detektion und einer Top-Down-Klassifizierung, wie es in Abschnitt 1.2 motiviert wurde. Bezüglich der Bottom-Up-Detektion unterscheidet sich das Verfahren noch stark von dem in Kapitel 6 beschriebenen Vorgehen. Thematische Überschneidungen bestehen insbesondere in der Top-Down-Stufe, die in den wesentlichen Punkten mit den in Abschnitt 6.5 beschriebenen Methoden übereinstimmt.

In [124] wird ein Verfahren vorgeschlagen, das die Erkennung von Proto-Objekten und Proto-Szenen miteinander verknüpft. Die Beiträge des zweitgenannten Autors umfassen die Methodik zur Szenenerkennung, die nicht in dieser Arbeit thematisiert wurde, sowie eine Erweiterung am verwendeten Bag-Of-Features-Modell. Letztere geht auf eine vorausgegangene Publikation des selben Autors zurück (Grzeszick et al. [50]), die in Abschnitt 3.4.2 thematisiert wurde. Bei der Methodik zur Detektion der Proto-Objekte handelt es sich um einen Beitrag des hiesigen Autors. Dieser umfasst insbesondere die thematischen Überschneidungen zur Segmentierung und Salienzdetektion aus Abschnitt 6.3 bzw. Abschnitt 6.4.

LITERATURVERZEICHNIS

[1] M. Jenkin, L. Harris, *Vision and Attention.* Springer, 2001.

[2] K. Grauman, B. Leibe, "Visual object recognition," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 5, no. 2, pp. 1–181, 2011.

[3] L. Elazary, L. Itti, "Interesting objects are visually salient," *Journal of Vision*, vol. 8, no. 3, pp. 1–15, 2008.

[4] D. Vernon, *Machine Vision: Automated Visual Inspection and Robot Vision.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1991.

[5] M. Casares, S. Velipasalar, "Light-weight salient foreground detection for embedded smart cameras," in *Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2008, pp. 1–7.

[6] C. Guo, L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[7] G. Zhang, Z. Yuan, N. Zheng, X. Sheng, T. Liu, "Visual saliency based object tracking," in *Proceedings of the 9th Asian conference on Computer Vision (ACCV).* Springer, 2010, vol. 2, pp. 193–203.

[8] S. Minut, S. Mahadevan, "A reinforcement learning model of selective visual attention," in *Proceedings of the 5th International Conference on Autonomous Agents (AGENTS).* New York, NY, USA: ACM, 2001, pp. 457–464.

[9] K. Shubina, J. K. Tsotsos, "Visual search for an object in a 3D environment using a mobile robot," *Computer Vision and Image Understanding*, vol. 114, no. 5, pp. 535–547, 2010.

[10] B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[11] R. C. Gonzalez, R. E. Woods, *Digital Image Processing (2nd Edition).* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2002.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[13] T. Lindeberg, "Scale-space theory: A basic tool for analysing structures at different scales," *Journal of Applied Statistics*, pp. 224–270, 1994.

[14] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 1999, pp. 1150–1157.

[15] L. Itti, C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[16] P. Oza, "Image segmentation in computer vision and review of various segmentation techniques," *International Journal of Computer Applications*, vol. NTSACT, no. 5, pp. 29–33, 2011.

[17] J. Freixenet, X. Munoz, D. Raba, J. Marti, X. Cufi, "Yet another survey on image segmentation: Region and boundary information integration," in *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, vol. 3, 2002, pp. 408–422.

[18] B. Peng, L. Zhang, D. Zhang, "A survey of graph theoretical approaches to image segmentation," *Pattern Recognition*, vol. 46, no. 3, pp. 1020–1038, 2013.

[19] P. F. Felzenszwalb, D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[20] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," in *Proceedings of the American Mathematical Society, 7*, 1956, pp. 48–50.

[21] C. Undeman, T. Lindeberg, "Fully automatic segmentation of MRI brain images using probabilistic anisotropic diffusion and multi-scale watersheds," in *Proceedings of the 4th International Conference on Scale Space Methods in Computer Vision (Scale Space)*. Springer, 2003, pp. 641–656.

[22] M. Drauschke, "An irregular pyramid for multi-scale analysis of objects and their parts," in *IAPR-TC-15 Workshop on Graph-based Representations in Pattern Recognition*, 2009, pp. 293–303.

[23] H. Niemann, "Klassifikation von mustern," http://www5.cs.fau.de/fileadmin/Persons/NiemannHeinrich/klassifikation-von-mustern/m00links.html, 2003.

[24] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 6, no. 2, pp. 559–572, 1901.

[25] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893 vol. 1.

[26] K. Mikolajczyk, C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[27] S. Geman, E. Bienenstock, R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[28] C. Cortes, V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

[29] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., 1995.

[30] L. Breiman, "Random forests," in *Machine Learning*, 2001, pp. 5–32.

[31] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*. Wadsworth, 1984.

[32] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[33] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[34] B. S. Everitt, S. Landau, M. Leese, *Cluster Analysis*, 4th ed. Wiley Publishing, 2009.

[35] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[36] R. Gray, "Vector quantization," *IEEE ASSP Magazine (Acoustics, Speech, and Signal Processing)*, vol. 1, no. 2, pp. 4–29, 1984.

[37] F. P. Preparata, M. I. Shamos, *Computational Geometry: An Introduction*. Springer, 1985.

[38] P. Perona, "Visual recognition circa 2008," in *Object Categorization: Computer and Human Perspectives*, S. J. Dickinson, Ed. Cambridge University Press, 2009.

[39] G. Kootstra, D. Kragic, "Fast and bottom-up object detection, segmentation, and evaluation using gestalt principles," in *Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 3423–3428.

[40] A. Andreopoulos, J. K. Tsotsos, "50 years of object recognition: Directions forward," *Computer Vision and Image Understanding*, vol. 117, no. 8, pp. 827 – 891, 2013.

[41] L. Wiskott, J.-M. Fellous, N. Krüger, C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[42] A. Vedaldi, B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[43] K. Mikolajczyk, C. Schmid, "Indexing based on scale invariant interest points," in *Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV)*, vol. 1, 2001, pp. 525–531.

[44] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of the ECCV 2004 International Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1–22.

[45] G. Salton, M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986.

[46] M. Varma, A. Zisserman, "Texture classification: Are filter banks necessary?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2003, pp. 691–698.

[47] E. Nowak, F. Jurie, B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proceedings of the 9th European Conference on Computer Vision (ECCV)*, A. Leonardis, H. Bischof, A. Pinz, Eds. Springer, 2006, vol. 3954, pp. 490–503.

[48] K. Grauman, T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2005, pp. 1458–1465 Vol. 2.

[49] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 2169–2178.

[50] R. Grzeszick, L. Rothacker, G. Fink, "Bag-of-features representations using spatial visual vocabularies for object classification," in *Proceedings of the 20th IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 2867–2871.

[51] A. Torralba, K. P. Murphy, W. T. Freeman, M. A. Rubin, "Context-based vision system for place and object recognition," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2003, pp. 273–280.

[52] M. Osadchy, Y. L. Cun, M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *Journal of Machine Learning*, vol. 8, pp. 1197–1215, 2007.

[53] M. Mathieu, M. Henaff, Y. LeCun, "Fast training of convolutional networks through FFTs," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

[54] P. Viola, M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[55] A. Vedaldi, V. Gulshan, M. Varma, A. Zisserman, "Multiple kernels for object detection," in *Proceedings of the 12th International Conference on Computer Vision (ICCV)*, 2009.

[56] K. E. A. van de Sande, J. Uijlings, T. Gevers, A. Smeulders, "Segmentation as selective search for object recognition," in *Proceedings of the 13th IEEE International Conference on Computer Vision (ICCV)*, 2011.

[57] L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[58] W. James, *The Principles of Psychology.* H. Holt, 1918.

[59] A. H. Clarke, C. Steineke, H. Emanuel, "High image rate eye movement measurement," in *Bildverarbeitung für die Medizin 2000*, 2000, pp. 398–402.

[60] M. I. Posner, M. J. Nissen, W. C. Ogden, "Attended and unattended processing modes: the role of set for spatial location." *Modes of Perceiving and Processing Information*, pp. 137–157, 1978.

[61] A. Treisman, G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[62] H. Gleitman, J. Jonides, "The effect of set on categorization in visual search," *Perception & Psychophysics*, vol. 24, no. 4, pp. 361–368, 1978.

[63] C. Koch, S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[64] M. Mishkin, L. G. Ungerleider, "Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys," *Behavioural Brain Research*, vol. 6, no. 1, pp. 57–77, 1982.

[65] J. Raymond, K. Shapiro, K. M. Arnell, "Temporary suppression of visual processing in an RSVP task: an attentional blink?" *Journal of experimental psychology. Human perception and performance*, vol. 18, no. 3, pp. 849–860, 1992.

[66] M. M. Chun, M. C. Potter, "A two-stage model for multiple target detection in rapid serial visual presentation," *Journal of Experimental Psychology: Human Perception and Performance*, pp. 109–127, 1995.

[67] R. A. Rensink, J. K. O'Regan, J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," in *Psychological Science*, vol. 8, no. 5, 1997, pp. 368–373.

[68] R. Desimone, J. Duncan, "Neural mechanisms of selective visual attention," *Annual Review of Neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[69] S. P. Vecera, M. J. Farah, "Does visual attention select objects or locations?" *Journal of Experimental Psychology*, vol. 123, no. 2, pp. 146–160, 1994.

[70] J. Duncan, G. Humphreys, R. Ward, "Competitive brain activity in visual attention," *Current Opinion in Neurobiology*, vol. 7 (2), pp. 255–261, 1997.

[71] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[72] G. Hartmann, *Gestalt Psychology: A Survey of Facts and Principles.* Ronald Press Company, 1935.

[73] R. Kimchi, Y. Yeshurun, A. Cohen-Savransky, "Automatic, stimulus-driven attentional capture by objecthood," *Psychonomic Bulletin & Review*, vol. 14, no. 1, pp. 166–172, 2007.

[74] R. Desimone, "Visual attention mediated by biased competition in extrastriate visual cortex," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 353, no. 1373, pp. 1245–1255, 1998.

[75] I. Rock, C. M. Linnett, P. Grant, A. Mack, "Perception without attention: Results of a new method," *Cognitive Psychology*, vol. 24, no. 4, pp. 502–534, 1992.

[76] D. J. Simons, C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events," *Perception*, vol. 28, pp. 1059–1074, 1999.

[77] J. Spinks, J. Zhang, P. Fox, J. Gao, L. Hai Tan, "More workload on the central executive of working memory, less attention capture by novel visual distractors: evidence from an fMRI study," *Neuroimage*, vol. 23, pp. 517–524, 2004.

[78] D. Simons, "Attentional capture and inattentional blindness," *Trends in Cognitive Science*, vol. 4, no. 4, pp. 147–155, 2000.

[79] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, 2006.

[80] A. Borji, L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[81] P.-E. Forssén, D. Meger, K. Lai, S. Helmer, J. J. Little, D. G. Lowe, "Informed visual search: Combining attention and object recognition," in *Proceedings of the 2008 IEEE International Conference on Robotics and Automation*, 2008, pp. 935–942.

[82] L. Itti, C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489–1506, 2000.

[83] J. Harel, C. Koch, P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, T. Hoffman, Eds. MIT Press, 2007, pp. 545–552.

[84] X. Hou, L. Zhang, "Saliency detection: A spectral residual approach," in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[85] D. Gao, V. Mahadevan, N. Vasconcelos, "On the plausibility of the discriminant centersurround hypothesis for visual saliency," *Journal of Vision*, pp. 1–18, 2008.

[86] R. Achanta, S. Hemami, F. Estrada, S. Süsstrunk, "Frequency-tuned Salient Region Detection," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1597 – 1604.

[87] Y. Zhai, M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proceedings of the 14th ACM International Conference on Multimedia*. ACM, 2006, pp. 815–824.

[88] J. Li, M. Levine, X. An, X. Xu, H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 996–1010, 2013.

[89] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, C. Koch, "Attentional selection for object recognition: A gentle way," in *Proceedings of the 2nd International Workshop on Biologically Motivated Computer Vision (BMCV)*. London, UK, UK: Springer, 2002, pp. 472–479.

[90] U. Rutishauser, D. Walther, C. Koch, P. Perona, "Is bottom-up attention useful for object recognition?" in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2004, pp. II–37–II–44 Vol.2.

[91] D. Walther, C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[92] D. Comaniciu, P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[93] H. Jo, A. Ojha, M. Lee, "A study on region of interest of a selective attention based on gestalt principles," in *Neural Information Processing*, M. Lee, A. Hirose, Z.-G. Hou, R. Kil, Eds. Springer, 2013, vol. 8228, pp. 41–48.

[94] J. Wu, L. Zhang, "Gestalt saliency: Salient region detection based on gestalt principles," in *Proceedings of the 20th IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 181–185.

[95] B. Alexe, T. Deselaers, V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[96] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, S.-M. Hu, "Global contrast based salient region detection," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 409–416.

[97] V. Navalpakkam, L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 2049–2056.

[98] Q. Zhao, C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of vision*, vol. 11, no. 3, 2011.

[99] A. Andreopoulos, J. K. Tsotsos, "On sensor bias in experimental methods for comparing interest-point, saliency, and recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 110–126, 2012.

[100] T. Judd, K. Ehinger, F. Durand, A. Torralba, "Learning to predict where humans look," in *Proceedings of the 12th International Conference on Computer Vision (ICCV)*, 2009.

[101] R. Lin, W. Lin, "A computational visual saliency model based on statistics and machine learning," *Journal of Vision*, vol. 14, no. 5, 2014.

[102] A. Oliva, A. Torralba, M. S. Castelhano, J. M. Henderson, "Top-down control of visual attention in object detection," in *Proceedings of the 2003 IEEE International Conference on Image Processing (ICIP)*, 2003, pp. 253–256.

[103] F. Miau, L. Itti, "A neural model combining attentional orienting to object recognition: preliminary explorations on the interplay between where and what," in *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 1, 2001, pp. 789–792 vol.1.

[104] I. Kokkinos, P. Maragos, A. Yuille, "Bottom-up and top-down object detection using primal sketch features and graphical models," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1893–1900.

[105] T. Avraham, M. Lindenbaum, "Attention-based dynamic visual search using inner-scene similarity: Algorithms and bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 251–264, 2006.

[106] H. Tagare, K. Toyama, J. Wang, "A maximum-likelihood strategy for directing attention during visual search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 5, pp. 490–500, 2001.

[107] C. Bandera, F. J. Vico, J. M. Bravo, M. E. Harmon, L. C. B. Iii, "Residual Q-learning applied to visual attention," in *Proceedings of the 13th International Conference on Machine Learning (ICML)*, 1996, pp. 20–27.

[108] L. Paletta, G. Fritz, C. Seifert, "Q-learning of sequential attention for visual object recognition from informative local descriptors," in *In Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005, pp. 649–656.

[109] ——, "Cascaded sequential attention for object recognition with informative local descriptors and Q-learning of grouping strategies," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) - Workshops*, 2005, pp. 94–94.

[110] S. Frintrop, G. Backer, E. Rome, "Goal-directed search with a top-down modulated computational attention system," in *Pattern Recognition*, W. Kropatsch, R. Sablatnig, A. Hanbury, Eds. Springer, 2005, vol. 3663, pp. 117–124.

[111] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*. Springer, 2006, vol. 3899.

[112] D. Meger, M. Muja, S. Helmer, A. Gupta, C. Gamroth, T. Hoffman, M. Baumann, T. Southey, P. Fazli, W. Wohlkinger, P. Viswanathan, J. Little, D. Lowe, J. Orwell, "Curious george: An integrated visual search platform," in *Proceedings of the 7th Canadian Conference on Computer and Robot Vision (CRV)*, 2010, pp. 107–114.

[113] J.-Y. Zhu, J. Wu, Y. Wei, E. Chang, Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3218–3225.

[114] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 862–875, 2015.

[115] Y. Rubner, C. Tomasi, L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[116] L. I. Rudin, S. Osher, E. Fatemi, "Nonlinear total variation based noise removal algorithms," in *Proceedings of the 11th Annual International Conference of the Center for Nonlinear Studies on Experimental Mathematics: Computational Issues in Nonlinear Science*, 1992, pp. 259–268.

[117] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[118] T. Judd, F. Durand, A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *Computer Science and Artificial Intelligence Lab (CSAIL)*, 2012, pp. TR–2012–001.

[119] A. Borji, L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *CVPR 2015 workshop on "Future of Datasets"*, 2015, arXiv preprint arXiv:1505.03581.

[120] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[121] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[122] ——, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html, 2012.

[123] F. Nasse, G. A. Fink, "A bottom-up approach for learning visual object detection models from unreliable sources," in *Proceedings of Pattern Recognition - Joint 34th DAGM and 36th OAGM Symposium*, 2012, pp. 488–497.

[124] F. Nasse, R. Grzeszick, G. A. Fink, "Toward object recognition with proto-objects and proto-scenes," in *Proceedings of the 9th International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, 2014, pp. 284–291.