Inhaltsanalyse unter Zuhilfenahme von Metadaten

Seminararbeit

am

Lehrstuhl für Praktische Informatik IV
Prof. Dr.-Ing. W. Effelsberg
Universität Mannheim

Eingereicht von
Alexander Bellm

Betreuer von
Prof. Dr. Heiner Stuckenschmidt

Entstanden im Rahmen des
Seminars „Bridging the Semantic Gap“
im SS 2006

Inhaltsverzeichnis

1. Einleitung 3
2. Objektgrenzenerkennung für ontologiegestützte Bildklassifizierung 4
    2.1 System 4
    2.2 Objekterkennung 5
        2.2.1 Randerkennung 5
        2.2.2 Bereichentwicklung 6
        2.2.3 Zusammenlegen benachbarter Bereiche 7
    2.3. Ergebnisse 9
3. Wissensunterstützte semantische Videoobjekterkennung 11
    3.1 System 11
    3.2 Multimediaanalyseontologie 12
    3.3 Bereichsspezifische Ontologie 13
    3.4 Ablauf 14
    3.5 Ergebnisse 15
    3.6 Geplante Systemerweiterungen 17
4. Systemvergleich und Fazit 17

1. Einleitung

Betrachtet ein Mensch ein Bild oder Video, so kann er gleich das Erkannte interpretieren und, sofern es ihm bekannt ist, auch benennen. Er erkennt ein Haus als Haus und ein Auto als Auto.
Ein Computer kann dies nicht. Daher versuchen viele Forschungseinrichtungen anhand einfacher Szenarien dem Computer Fähigkeiten dieser Art beizubringen. Ziel dabei ist es, das Wesentliche aus einem Bild oder Video zu erkennen und es anschließend zu benennen. Hierfür gibt es natürlich eine Vielzahl von Ansätzen. Eindeutig ist aber, dass Zusatzinformationen dabei helfen. Wenn zum Beispiel bekannt ist, dass das betrachtete Bild aus dem Bereich Sport stammt, lassen sich bestimmte Merkmale besser einordnen oder erkennen. Solche Zusatzinformationen können dem Computer beispielsweise als Metadaten oder Ontologien bereitgestellt werden.

In dieser Seminararbeit werden zwei Systeme aus Forschungsarbeiten vorgestellt und verglichen. Beide nutzen spezielle Zusatzangaben um eine Inhaltsanalyse bei Bildern beziehungsweise Videos durchzuführen.
System eins ist Forschungsarbeit an der Universität von Texas in Dallas, nennt sich „Objektgrenzenerkennung für ontologiegestützte Bildklassifizierung“ und versucht durch das Erkennen der wesentlichen Objekte des Bildes Schlagwörter zu generieren, die das Bild beschreiben.
Das zweite System ist Teil der Forschung an der Universität von Thessaloniki, heißt „Wissensunterstützte semantische Videoobjekterkennung“ und nutzt für diese Videoobjekterkennung das Wissen aus einer bereichsspezifischen Ontologie, sowie einer Ontologie, welche alle Faktoren der Multimediaanalyse beschreibt.
Ausgangspunkt beider Systeme sind die für den Computer direkt erkennbaren low-level Eigenschaften des Bildes oder Videos wie Farb-, Helligkeits- und Intensitätswerte.

2. Objektgrenzenerkennung für ontologiegestützte Bildklassifizierung

Wie in der Einleitung bereits erwähnt, stammt diese Forschungsarbeit aus der Universität von Texas, in Dallas. Im Folgenden wird das grundlegende System der Arbeit vorgestellt, sowie auf dessen Hauptaufgaben und einige Beispiele des Leistungsstandes eingegangen.

2.1 System

Das System nutzt sowohl Ontologien als auch Neuronale Netze zur Objekterkennung um ein hohes Level an Präzision bei der automatischen inhaltsspezifischen Bildklassifizierung zu erreichen.

Die hierbei verwendete Ontologie wird mit Hilfe von Beispielbildern von Hand erzeugt. Inhalt der Ontologie sind Schlagwörter in Zusammenhang mit für die Schlagwörter charakteristischen Eigenschaften. Je detaillierter die Ontologie ausgearbeitet ist, desto besser sind am Ende die Ergebnisse des Systems.

Um nun ein Bild zu klassifizieren, wird als erstes eine Objekterkennung durchgeführt. Auf das Erkennen dieser Objekte wird im weiteren Verlauf genauer eingegangen. Die erkannten Objekte werden durch ein neuronales Netzwerk identifiziert. Diese Identifizierung wird von Durchgang zu Durchgang besser, da das neuronale

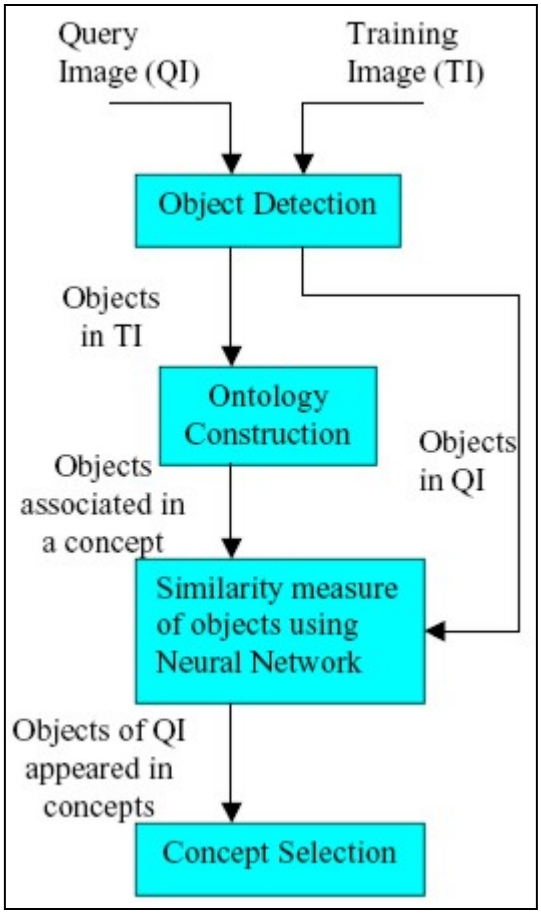


Abb. 1: Ablaufdiagramm des Systems

Netzwerk dazulernt. Auf die neuronalen Netzte wird nicht näher eingegangen. Sie werden als Black-Box betrachtet, welche die gewünschten Ausgaben liefert. Anschließend werden dem Bild anhand der identifizierten Objekte, mit Hilfe eines skalierbaren Algorithmus im „Concept-Selection-Modul“ unter Zuhilfenahme der Ontologie, Schlagwörter zugeteilt. Diese Schlagwörter sollen das Bild klassifizieren und charakterisieren.

2.2 Objekterkennung

Die in dieser Forschungsarbeit entwickelte Objekterkennung arbeitet in drei Schritten, der Randerkennung, der Bereichsentwicklung und abschließend dem Zusammenlegen benachbarter Bereiche.

2.2.1 Randerkennung

Grundlage der Randerkennung sind die Intensitätswerte jedes einzelnen Bildpixels. Ziel ist es alle Pixel in zwei Gruppen einzuteilen, die Gruppe der Randpixel EPS (edge pixel set) und die Gruppe der Bereichspixel RPS (region pixel set).
Dieser Einteilung werden vier Szenarien zugrunde gelegt, je eines für jede mögliche Richtung des Randes.

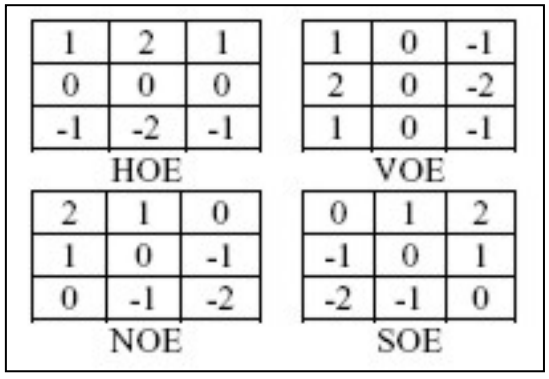


Abb. 2: Definition der Randrichtungen

Jedes Pixel (x, y) wird also mit seinen acht Nachbarn mit einer bestimmten Gewichtung verglichen.

(x-1, y-1)	(x-1, y)	(x-1,y+1)
(x, y-1)	(x, y)	(x, y+1)
(x+1,y-1)	(x+1,y)	(x+1,y+1)

Abb. 3: Definition der Koordinaten

Anhand einer Formel wird dann entschieden welcher Gruppe der Pixel zugeteilt wird.

$$
\begin{aligned}
HOE(x, y)_i &= | I(x-1,y-1) + 2I(x,y-1) + I(x+1,y-1) \\
&\quad - I(x-1,y+1) - 2I(x,y+1) - I(x+1,y+1) | \\
VOE(x, y)_i &= | I(x-1,y-1) + 2I(x-1,y) + I(x-1,y+1) \\
&\quad - I(x+1,y-1) - 2I(x+1,y) - I(x+1,y+1) | \\
NOE(x, y)_i &= | I(x,y-1) + 2I(x-1,y-1) + I(x-1,y) \\
&\quad - I(x+1,y) - 2I(x+1,y+1) - I(x,y+1) | \\
SOE(x, y)_i &= | I(x,y-1) + 2I(x+1,y-1) + I(x+1,y) \\
&\quad - I(x-1,y) - 2I(x-1,y+1) - I(x,y+1) | \\
MOE(x, y)_i &= \max \{HOE(x, y)_i , VOE(x, y)_i, \\
&\quad NOE(x, y)_i, SOE(x, y)_i \}
\end{aligned}
$$

Abb. 4: Die schrittweise Berechnung von (MOE)

Dabei wird der abschließende Wert der Berechnung (MOE) mit einem anpassbaren Grenzwert verglichen. Ist MOE größer als dieser Grenzwert, wird der Pixel der Gruppe der Randpixel zugeteilt.

2.2.2 Bereichentwicklung

Das Generieren von einzelnen Bereichen bzw. Regionen aus den zwei Pixelmengen wird in einer Schleife durchgeführt:

- ⇒ Ein Pixel wird aus der Menge der Bereichspixel (RPS) zufällig ausgewählt. Dieser Pixel wird aus RPS gestrichen und einer neuen nummerierten Region zugeteilt. Weil er einziger Pixel dieser Region ist, ist er automatisch Grenzpixel (boundary pixel).
  Ein Pixel ist Grenzpixel wenn mindestens einer seiner acht Nachbarpixel nicht zur gleichen Region gehört. Anderenfalls wäre er innerer Pixel (inner pixel).
- ⇒ Auswahl der Nachbarknoten, die Bereichspixel, also in RPS sind.

⇒ Überprüfung dieser Pixel auf folgende Eigenschaften:

- Die Farbabweichung zwischen dem Pixel und seinen Nachbarpixel in der Region sind vernachlässigbar klein
- Die Farbabweichung zwischen dem Pixel und dem Durchschnittswert der Nachbarpixel in der Region sind vernachlässigbar klein
- Die Farbabweichung zwischen dem Pixel und dem Durchschnittswert der Region sind vernachlässigbar klein sind diese Bedingungen erfüllt, wird der entsprechende Pixel der Menge der Bereichspixel entnommen und dieser Region zugeteilt.

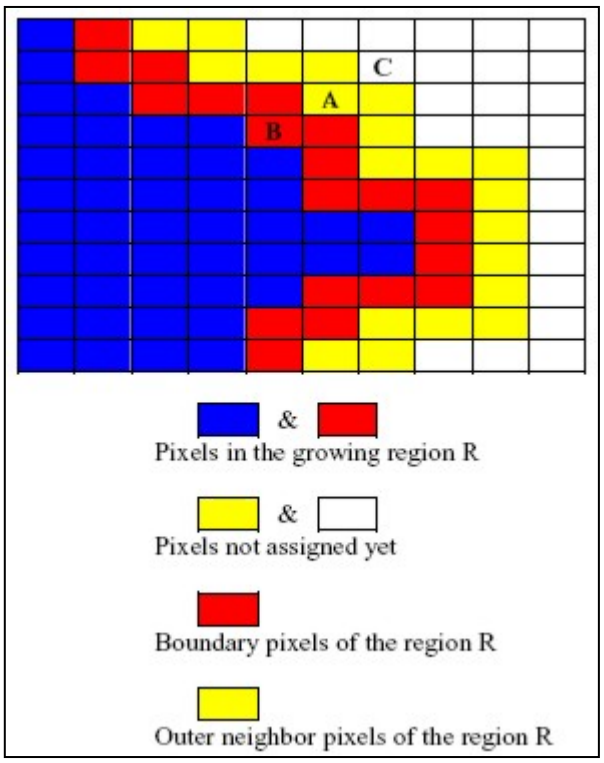


Abb. 5: Regionsentwicklung

⇒ Die Einteilung in Grenzpixel und inneren Pixel wird durchgeführt.

⇒ Auswahl der Nachbarknoten aller Pixel dieser Region, die noch in RPS sind.

⇒ Weiteres Durchlaufen der Schleife, bis die Region nicht mehr wächst.

⇒ Wenn RPS noch Pixel enthält, erfolgt erneut eine zufällige Wahl eines Pixels und die Zuteilung zu einer neuen Region.

⇒ Die Schleife wird von neuem durchlaufen, bis RPS keine Pixel mehr enthält.

2.2.3 Zusammenlegen benachbarter Bereiche

Probleme, die bei der Bereichsentwicklung auftreten können, sind so genannte „noise regions“ und das ein Objekt fälschlicherweise in mehrere Regionen eingeteilt wurde. „Noise regions“ sind kleine Regionen bzw. Pixelmengen oder einzelne Pixel,

von denen angenommen wird, dass sie vernachlässigbar sind. Sie werden nicht zur Klassifizierung des Bildes benötigt.

Um diese Probleme zu lösen wird die Matrix Oriented Tequnique (MOT) verwendet.

Bei diesem Verfahren wird eine Matrix aufgestellt, die durch jeden Eintrag ein Bildpixel repräsentiert. Der Wert des Eintrages ist die Regionsnummer der Region, der der Pixel zugeteilt wurde. Randpixel erhalten den Wert -1. Mit dieser Matrix kann nun bestimmt werden, welche Regionen benachbart sind. Hierzu geht man die Matrix Zeile für Zeile und anschließend Spalte für Spalte durch. Sind nun zwischen zwei Bereichspixel verschiedener Regionen nur -1er in einer Zeile oder Spalte, werden diese als benachbart angenommen. Wie in der zweiten Abbildung zu sehen ist, können dabei aber falsche Regionen als benachbart angenommen werden. Hier sind in der 5. Zeile nur -1er zwischen Region 2 und 3, sie sind aber dennoch nicht benachbart. Um diesen Effekt zu vermeiden wird ein Maximalwert definiert, der angibt wie viele solcher -1er zwischen zwei Regionen maximal sein dürfen, damit sie als benachbart angenommen werden können. Wenn man also weiß welche Regionen benachbart sind, kann man die oben genannten Probleme beheben.

-1	-1	5	5	5	5	-1	3
2	-1	5	5	5	5	-1	3
2	-1	-1	5	5	-1	-1	3
2	2	-1	5	5	-1	3	3
2	2	-1	5	5	-1	3	3
2	2	2	-1	-1	3	3	3
2	2	2	2	2	-1	-1	3
2	2	2	2	2	2	-1	3
2	2	2	2	2	2	2	-1

2	-1	5	5	5	5	5	-1
2	-1	5	5	5	5	-1	-1
2	-1	5	5	5	5	-1	3
2	-1	5	5	5	5	-1	3
2	-1	-1	-1	-1	-1	-1	3
2	-1	4	4	4	4	-1	3
2	-1	4	4	4	4	-1	3
2	-1	4	4	4	4	4	-1
2	-1	4	4	4	4	4	-1

Abb. 6: Beispielmatrizen für MOT

Tritt nun eine „noise region“ auf, also eine Region die kleiner als ein definierter Grenzwert ist, wird sie der Nachbarregion zugeordnet, welche die kleinste Farbabweichung aufweist. So werden unerwünschte, sehr kleine Regionen eliminiert.

Beim zweiten Problem, dass ein Objekt in verschieden Regionen aufgeteilt wurde, kann nur unter bestimmten Umständen behoben werden.

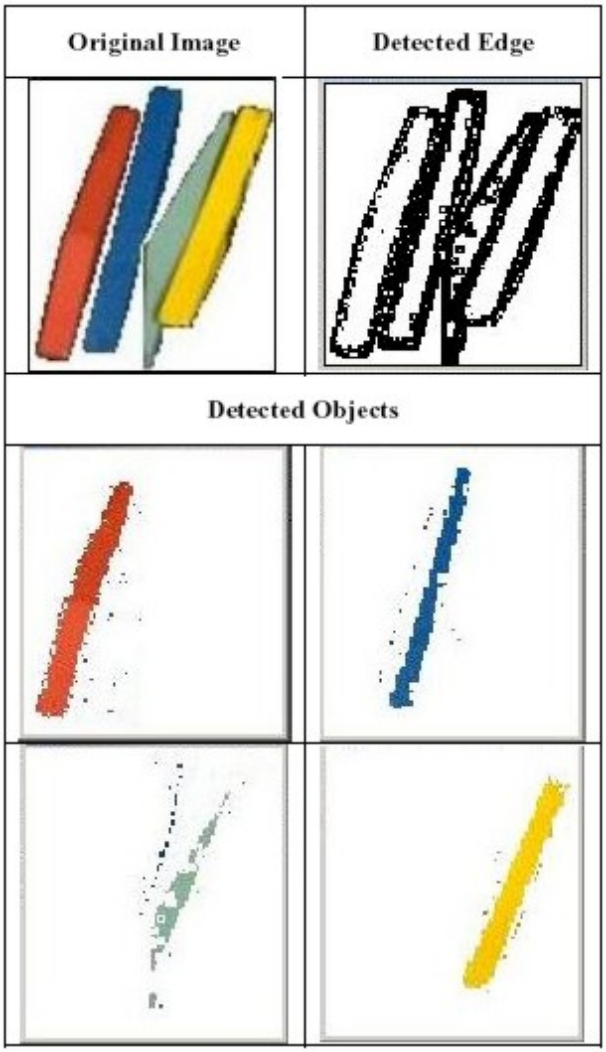
Hierzu wird jedem Randpixel ein Wert zugeteilt, der den Farbunterschied der zwei angrenzenden Regionen angibt. Ist

Abb. 7: Beispielbild

dieser Wert nun kleiner als ein variabler Grenzwert, so werden diese Regionen zusammengelegt. Dieser Grenzwert muss durch Testen ermittelt oder verfeinert werden, um das gewünschte Ergebnis zu erzielen. Durch dieses Verfahren können zum Beispiel die verschiedenen Flächen eines Tennis- oder Basketballs zu einer Region zusammengelegt werden.

2.3. Ergebnisse

Einen kleinen Überblick über die aktuelle Leistungsfähigkeit des Systems zeigen die nachfolgenden Beispielbilder, die mit diesem System bearbeitet wurden und deren Ergebnisse:



(a)

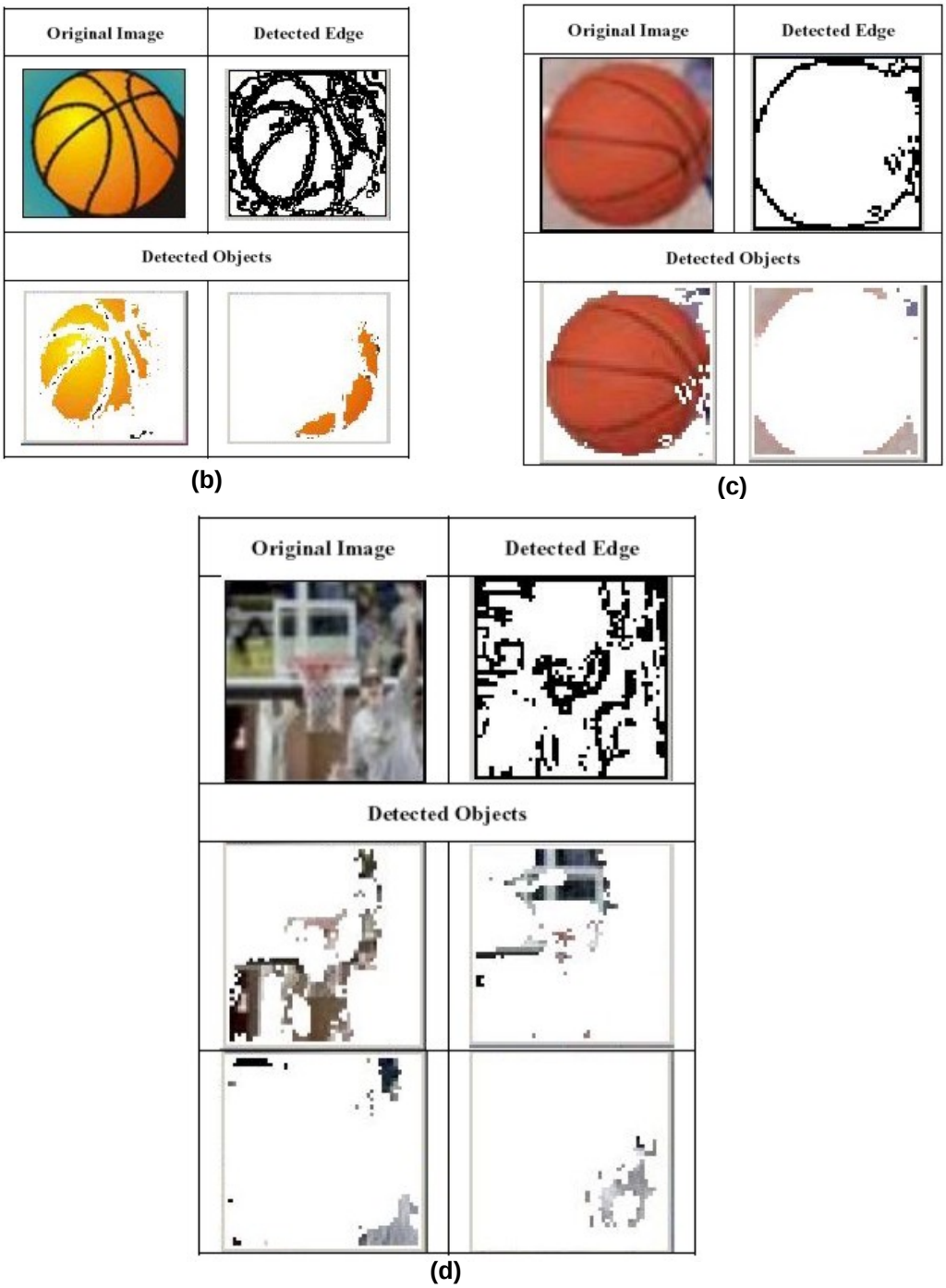


Abb. 8: Ergebnisse der Segmentierung

Bei diesen Beispielen wird eindeutig sichtbar, dass die hier verwendete Objekterkennung noch lange nicht ausgereift ist und nur sehr einfache Objekte mit möglichst gleichmäßiger Farbe zuverlässig erkennt. Beim letzten Beispiel, dem Basketballspiel, wird besonders deutlich, dass hier noch absolut keine Chance besteht, einzelne Objekte aus dem bild zu identifizieren.

3. Wissensunterstützte semantische Videoobjekterkennung

Diese Forschungsarbeit von der Universität von Thessaloniki beschäftigt sich mit der Objekterkennung in Videos. Auch hier wird das entwickelte System vorgestellt und auf dessen zugrunde liegenden Ontologien eingegangen. Zusätzlich werden ein paar Beispiele zum aktuellen Leistungstand gegeben.

3.1 System

Dieses System basiert auf zwei grundverschiedenen Ontologien, der Multimediaanalyseontologie (multimedia analysis ontology) und einer bereichsspezifischen Ontologie (domain specific ontologie).

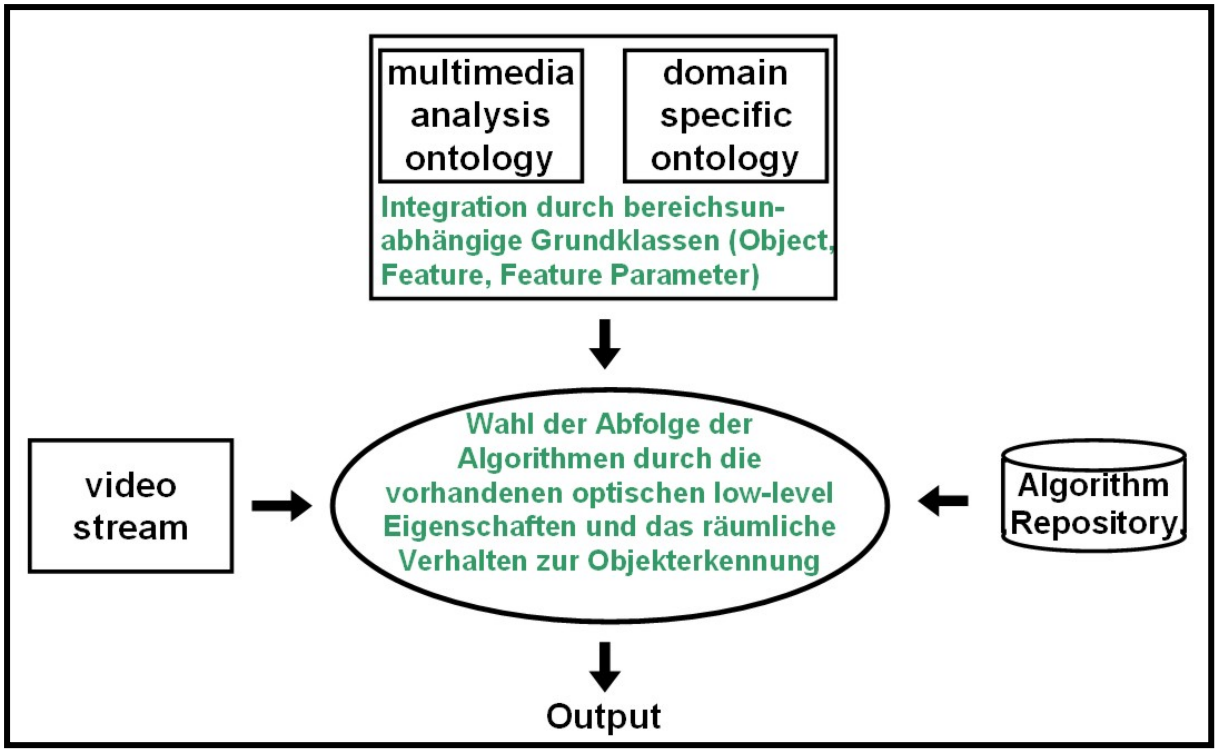


Abb. 9: Systemgrundriss

Diese zwei Ontologien werden durch bereichsunabhängige Grundklassen, welche in beiden vorkommen, verknüpft und dienen dann dem System zusammen mit den

low-level Eigenschaften des Videostreams als Entscheidungsgrundlage. Mit ihrer Hilfe entscheidet das System, welche low-level Algorithmen aus dem Algorithmenspeicher (Algorithm Repository) in welcher Reihenfolge auf den Videostream angewendet werden, um den gewünschten Output zu generieren.

3.2 Multimediaanalyseontologie

Diese Ontologie, wie auch die später genauer erläuterte bereichsspezifische Ontologie, wurde von externen Ontologieexperten entworfen.

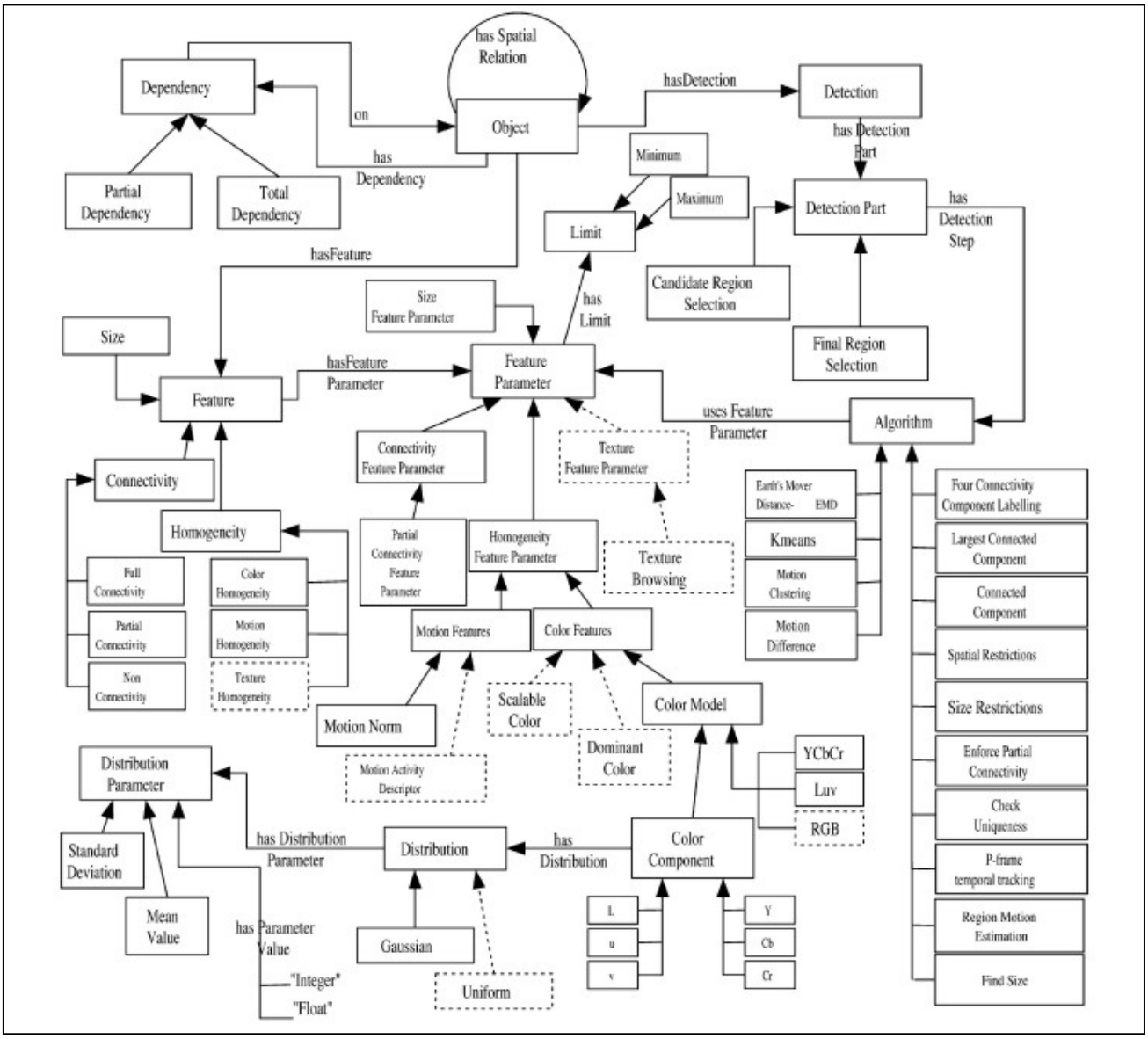
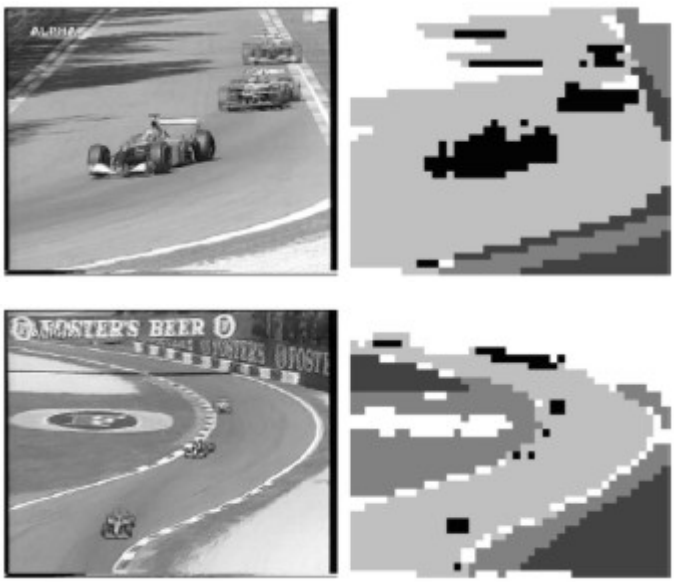


Abb. 10: Die Multimediaanalyseontologie

Die Multimediaanalyseontologie unterstützt die Erkennung von bereichsspezifischen Objekten und speichert alle ihre für das System relevanten Eigenschaften.
Die Wurzel der Ontologie ist die Superklasse „Object“. Alle vom System erkannten Videoobjekte sind dann Objekte der Klasse „Object“. In dem Teilbaum „Dependency“ werden zum Beispiel die Abhängigkeiten zu anderen Objekten angegeben. Die Superklasse „Feature“ beinhalten in ihren Unterklassen alle low-level Eigenschaften des Objekts wie die Größe alle Farbwerte und viele mehr. Räumliche Beziehungen zwischen den Objekten werden direkt in der Beziehung „has Spatial Relation“ angegeben.
Auch die angewendeten Algorithmen haben ihren Platz in der Ontologie und beziehen ihre benötigten Informationen ausschließlich aus dieser.

3.3 Bereichsspezifische Ontologie

Die in dem System verwendeten bereichsspezifischen Ontologien sind ziemlich klein und einfach und beschreiben nur einen sehr kleinen, eingeschränkten Bereich.
Als Beispiel für eine bereichsspezifische Ontologie dient hier die „Formula One domain ontology“. Es wird später aber noch kurz anhand anderer Beispiele auf andere Eigenschaften anderer Ontologien eingegangen.



Bei der „Formula One domain ontology“ wird das Szenario auf vier Objekte vereinfacht. Hierbei handelt es sich um Auto, Straße, Gras und Sand. Alle anderen Objekte, die keinem dieser vier zugeteilt werden können, werden von dem System später ignoriert. Diese vier Objekte haben folgende charakteristische Eigenschaften:

Abb. 11: Beispielsegmentierung Formel 1

Auto:

- zusammenhängende Region
- homogene Bewegung
- innerhalb des Objekts Straße
- Bewegungswert > Minimaler Wert
- Größe hat keinen Maximalwert

Gras:

- homogene Farbe
- teilweise zusammenhängende Region
- jede Teilregion > minimale Größe
- adjacent-to-Beziehung zu Straße
- adjacent-to-Beziehung zu Sand

Straße:

- zusammenhängende Region
- homogene Farbe
- Größe > vordefinierter Minimalwert
- größte Region des Videos

Sand:

- homogene Farbe
- teilweise zusammenhängende Region
- jede Teilregion > minimale Größe
- adjacent-to-Beziehung zu Straße
- adjacent-to-Beziehung zu Gras

Bei dieser recht kleinen Ontologie ist es sehr wichtig so genau wie möglich zu sein, um am Ende eine möglichst gute Videoobjekterkennung zu erreichen.
Durch diese bereichsspezifische Abgrenzung der bereichsspezifischen Ontologie, ist es sehr leicht, das System für andere Bereiche zu benutzen, da jeweils nur eine andere bereichsspezifische Ontologie benötigt wird. Alles andere bleibt aber identisch.

3.4 Ablauf

Zusammenfassend wird der Ablauf des Systems nochmals durch die wichtigsten Schritte beschrieben:

⇒ Input: Videostream
⇒ Segmentierung durch eine Farbwerttabelle (aufgestellt durch Training-Sets)
⇒ Automatische Objekterkennung durch die Multimediaanalyseontologie und das Anwenden von low-level Algorithmen
(Zwischenergebnis: Objekte mit ihren low-level Eigenschaften)

⇒ Benennung der relevanten Objekte durch das Vergleichen der vorhandene low-level Eigenschaften der Objekte, mit denen in der bereichsspezifischen Ontologie

⇒ Nichterkannte Objekte werden ignoriert

⇒ Output: Videostream mit den erkannten Objekten

Hierbei wird deutlich sichtbar, dass dieser Ansatz hauptsächlich auf visuellen Objekteigenschaften und auf den räumlichen Beziehungen der einzelnen Objekte zueinander basiert.

3.5 Ergebnisse

Im Folgenden werden ein paar Beispiele und Auswertungen aus unterschiedlichen Bereichen vorgestellt. Zuerst Beispiele aus dem Bereich der Formel 1.

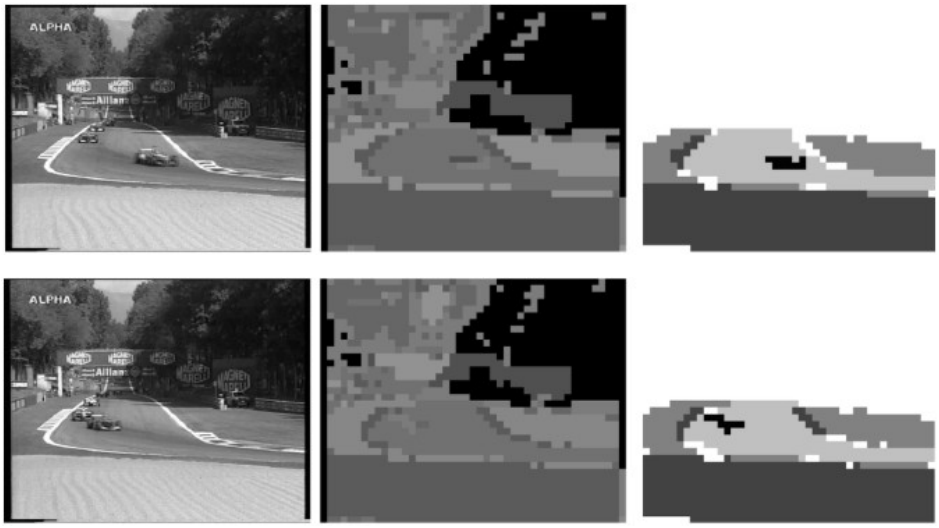


Abb. 12: Eingabevidoestream, Segmentierung und Resultat

Object	correct detections	false detections	missed
Road	97%	2%	1%
Grass	90%	7%	3%
Sand	88%	8%	4%
Car	74%	22%	7%

Abb. 13: Numerische Auswertung im Bereich Formel 1

Auch im Bereich des Fußballs wurde das System der „Wissensunterstützten semantischen Videoobjekterkennung“ getestet.

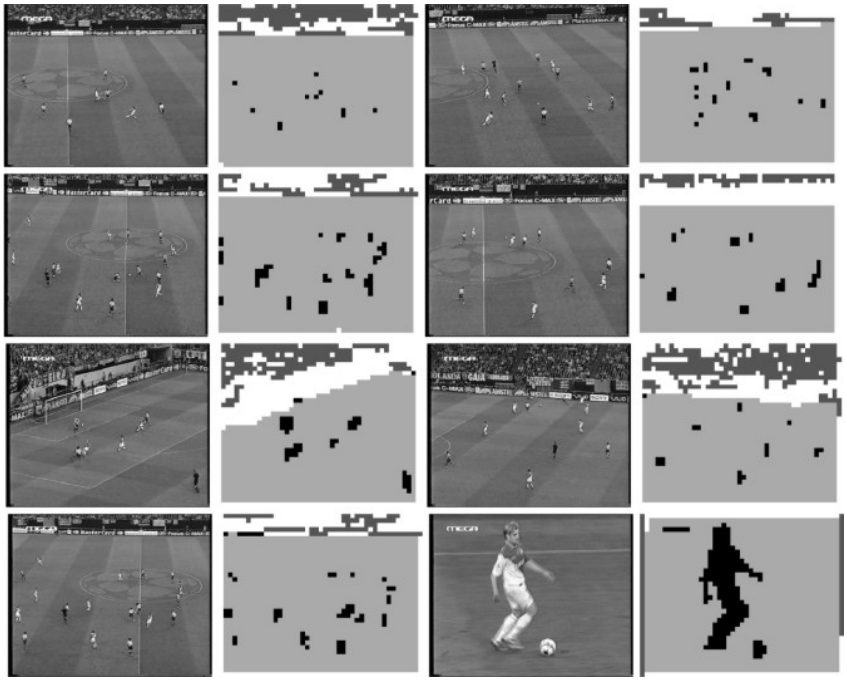


Abb. 14: Ergebnisse aus den Bereich Fußball

Object	correct detections	false detections	missed
Field	100%	0%	0%
Player	82%	5%	13%
Spectators	70%	2%	28%

Abb. 15: Numerische Auswertung im Bereich Fußball

Die Auswertungen zeigen in beiden Fällen durchaus akzeptable bis erstaunliche Werte.

Abschließend die Ergebnisse des Systems bei der Anwendung auf Strandvideos:

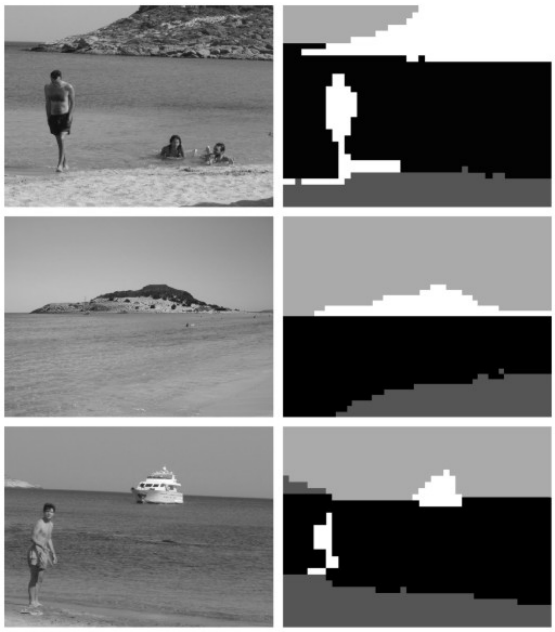


Abb. 16: Ergebnisse aus den Bereich der Strandvideos

Hierbei ist interessant, dass die räumlichen Beziehungen zwischen den Objekten die dominante Rolle einnehmen. Die bereichsspezifische Ontologie beschreibt Beziehungen wie:

- Sand is adjacent-to Sea
- Sea is below-of Sky
- Sky is above-of Sea
- Sky is above-of Sand

3.6 Geplante Systemerweiterungen

Das derzeitige Ziel der Forscher ist es, durch die Erweiterung der bereichsspezifischen Ontologien, eine komplexere Objektdarstellung zu erreichen. Hierzu wird versucht mehr Objekte, mehr low-level Eigenschaften der Objekte und mehr räumliche Beziehungen der Objekte zueinander in den bereichsspezifischen Ontologien aufzunehmen.
Außerdem wird daran gearbeitet in Zukunft auch Raum-Zeit Beziehungen im System zu berücksichtigen, sowohl zur Erkennung von Objekten als auch zur Erkennung von Events.

4. Systemvergleich und Fazit

Die beiden vorgestellten Systeme bedienen sich beide an Zusatzinformationen die anhand von Ontologien bereitgestellt werden. Trotzdem kann man sie nur schlecht

vergleichen, da bei beiden die Ontologien an völlig verschiedenen Stellen und mit ganz anderen Aufgaben zum Einsatz kommen.
Beim ersten System, der „Objektgrenzenerkennung für ontologiegestützte Bildklassifizierung“ wird zuerst eine Objekterkennung ohne Zuhilfenahme der Ontologie durchgeführt. Erst nach der Objekterkennung werden die Objekte mit Hilfe der Ontologie identifiziert und benannt.
Beim zweiten System, der „Wissensunterstützten semantischen Videoobjekterkennung“, werden hingegen die zwei Ontologien direkt zur Erkennung eingesetzt. Ihr Inhalt ist wichtige Wissensbasis für das System.

Bei beiden Systemen kann man sagen, dass sie erste zufrieden stellende Ergebnisse liefern, aber noch deutlich in ihren Kinderschuhen stecken.
In beiden Fällen wird schnell deutlich, dass erst eine kleine Menge an einfachen Bildern und Videos mit dem gewünschten Ergebnis bearbeitet werden können.
Das erste System arbeitet nur dann mit dem erwünschten Erfolg, wenn die Bilder nicht zu kompliziert sind, also die Anzahl der Objekte nicht zu groß wird, die zu erkennenden Objekte alle eine bestimmte Mindestgröße haben und die Objekte eine einheitliche Farbe haben, oder die Abweichung einen bestimmten kleinen Grenzwert nicht übersteigt. Auch ist es von Nöten, dass eine bestimmte Anzahl an Trainingsbildern bearbeitet wurde, um die Ontologie ausreichend aufgebaut zu haben und so die gewünschte Qualität der Erkennung zu garantieren.
Das zweite System setzt voraus, dass man den Bereich des Inhaltes des Videos kennt und eine entsprechende bereichsspezifische Ontologie zur Verfügung steht. Das System verlangt auch noch eine starke Vereinfachung des Themenbereiches und ist auf eine kleine Anzahl an Objekten beschränkt. Zudem müssen zu Beginn in einer Trainingsphase die Farbwerte für die einzelnen Objekte ermittelt werden.

Abschließend kann man also festhalten, dass die Systeme noch deutlich erweitert und verfeinert werden müssen, um einen allgemeinen praktischen Nutzen zu erhalten.