MPEG-7: Ein Standard zur Beschreibung von Multimedia-Inhalten

Marcel Richter
mrichter@techfak.uni-bielefeld.de

1. Juli 2003

Zusammenfassung

Dieser Text enstand im Rahmen des Seminars Bilddatenbanken im SS 2003 an der Technischen Fakultät/Universität Bielefeld.

Zunächst gibt der Text einen Überblick über den MPEG-7 Standard, der zur Beschreibung von Multimedia-Inhalten entwickelt wurde. Im zweiten Teil der Ausarbeitung wird detaillierter auf die Entwicklungen der Arbeitsgruppe "MPEG-7 Visual" eingegangen. Zum Schluss werden einige Anwendungen vorgestellt, die das MPEG-7 Format benutzen.

Inhaltsverzeichnis

1 Motivation

Auf Grund der großen Menge von Multimedia-Daten, die im Internet, in Werbefirmen, in Rundfunksendern aber auch auf dem privaten Computer vorliegen, besteht in vielen Bereichen die Notwendigkeit, diesen Bestand möglichst schnell und effizient nach gewünschten Daten zu durchsuchen. Aus vorherigen Vorträgen wurde deutlich, dass es sinnvoll ist, die Inhaltsbasierte Bildsuche der Textbasierten Suche vorzuziehen, da die nicht-automatische Annotation von Bildern sehr arbeitsaufwendig und daher meist teuer und zugleich auch noch fehlerbehaftet ist. Die MPEG (Moving Picture Experts Group) entwickelte einen Standard namens MPEG-7, der ein Format zur Inhaltsbeschreibung von audiovisuellen Medien darstellt.

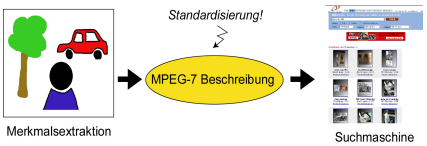

Abbildung 1: Arbeitsbereich des MPEG-7 Standards

Ausdrücklich soll hier betont werden, dass MPEG-7 kein Standard zur automatischen oder semi-automatischen Merkmalsextraktion ist, er stellt lediglich das Format bereit, welches ein externes Programm zur Speicherung der extrahierten Merkmale verwenden kann. Auch die Anwendungen, die auf den Inhaltsbeschreibungen arbeiten, wie z.B. Suchmaschinen, werden von MPEG-7 nicht spezifiziert. Hier sollen durch Wettbewerb zwischen den einzelnen Anbietern die bestmöglichen Resultate erzielt werden.

Warum aber ist ein Standard wie MPEG-7 nötig? Durch den Standard wird eine Kompatibilität an der Schnittstelle zwischen Merkmalsextraktion und den Anwendungen hergestellt. Außerdem bietet MPEG-7 die weltweit größte Menge an audiovisuellen Beschreibungsmöglichkeiten. Glaubt man der MPEG, dann werden alle Bereiche, die mit Multimedia in Kontakt sind, von dem MPEG-7 Standard profitieren.

Besonders sind dabei z.B. Digitale Multimedia Bibliotheken zu nennen, die im Rundfunk oder der Werbebranche eine Möglichkeit darstellen, um passende Inhalte zu einer bestimmten Suchanfrage zu finden. Mit MPEG-7 als Speicherformat könnten diese Suchanfragen aus einem gesprochenen Text, einer Skizze, Beispielbildern oder auch einer gesummten Melodie bestehen. Aber auch das Internet und der private PC werden nach Multimediainhalten durchsuchbar.

2 Typische Anwendungen

Eine typische Anwendung, bei der eine Inhaltsbasierte Suche verwendet wird, ist z.B. vorhanden, wenn eine Skizze als Suchanfrage gegeben ist. Die Anwendung berechnet dann auf Grundlage der Skizze, eine Menge von Bildern aus der Datenbank, die der Skizze ähnlich sind. So können z.B. Bilder mit ähnlichen Farbverteilungen aber auch Markenlogos, die ähnlich der Skizze sind, gefunden werden. Eine weitere Anwendung im visuellen Bereich ist das Erkennen von prägnanten Bewegungsmustern innerhalb eines Videos. So kann z.B. bei einem Fußballspiel in Echtzeit ein Tor erkannt werden und direkt über den Mobilfunk die Videosequenz des Torschusses versendet werden. Eine ähnliche Anwendung ist die Einbindung in Überwachungssysteme, bei denen bestimmte Bewegungsmuster zu einem Alarm führen, z.B. im miltärischen Bereich die Überwachung von Lufträumen.

Im Audiobereich ist eine Suchanfrage durch summen oder pfeifen möglich("query-by-humming"). Da innerhalb von MPEG-7 die Möglichkeit besteht, von der Klangfarbe der Instrumente zu abstrahieren, kann so eine Melodie gefunden werden, die der gesummten Suchanfrage entspricht. Es besteht auch die Möglichkeit nach Stimmen zu suchen. Als Suchanfrage verwendet man dabei ein Stimmensample. Bei einer Anfrage mit einem Stimmensample von Pavarotti, erhält man beispielsweise alle von Pavarotti gesungenen Lieder, alle gegebenen Interviews aber auch Fotos und Videos auf denen er zu sehen ist.

3 Standards

MPEG-7 standardisiert folgendes:

- Eine Menge von Deskriptoren
- Eine Menge von Description Schemes
- Eine Sprache, die sowohl die Deskriptoren als auch die Description Schemes definiert: DDL (Description Definition Language)
- Eine Möglichkeit zur Kodierung der Beschreibungen

Eine Übersicht der Standards ist in Abbildung 2 zu sehen.

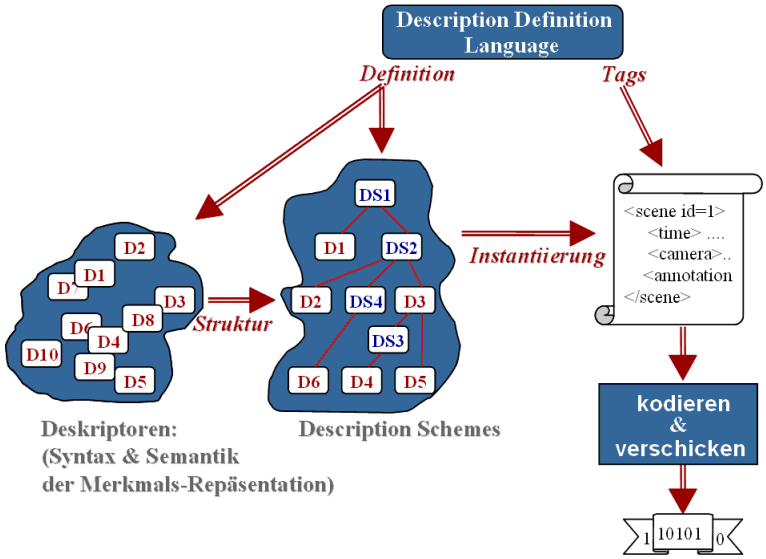

Abbildung 2: Standards von MPEG-7: Deskriptoren repräsentieren Merkmale des Inhalts, Description Schemes strukturieren Deskriptoren und Description Schemes und die Description Definition Language ist die Beschreibungssprache zur Definition von Deskriptoren und Description Schemes.

Deskriptoren

Ein Deskriptor ist eine Repräsentation eines Merkmals des Multimedia-Inhalts. Man kann in diesem Zusammenhang zwischen drei Gruppen von Merkmalen unterscheiden.

Die *katalogisierten Merkmale* beschreiben Meta-Informationen der Multimedia-Datei, wie z.B. Titel, Hersteller, Genre oder auch Zweck der Herstellung. Da diese Merkmale nicht automatisch aus dem Inhalt extrahiert werden können, muss man diese Daten weiterhin manuell annotieren oder falls vorhanden, aus einer dem Inhalt zugeordneten Header-Datei auslesen. Weiterhin enthalten die katalogisierten Merkmale Informationen über den Gebrauch des Materials, wie z.B. Urheberrechte und finanzielle Informationen. Auch das Speichermedium, das Kodierungsformat und eine eindeutige Identifikation gehören in diese Gruppe.

Die *semantischen Merkmale* geben Informationen über das "Wer, Was, Wann, Wo?" und liefern Informationen über Objekte und Ereignisse.

Die *strukturellen Merkmale* beschreiben den audiovisuellen Inhalt aus dem Blickwinkel seiner Struktur. Das Material wird in physische, räumliche und zeitliche Segmente geteilt und auf diesen Segmenten werden wiederum Beschreibungen der Signalbasierten Merkmale (Farbe, Textur, Form, Bewegung, Geräusche) vorgenommen.

Description Schemes

Ein Description Scheme spezifiziert die Struktur und die Semantik zwischen verschiedenen Deskriptoren und anderen Description Schemes.

Description Definition Language (DDL)

Sowohl die Deskriptoren als auch die Description Schemes werden in DDL definiert. DDL benutzt XML Schema Language als Grundlage für ihre Syntax, jedoch werden auch einige Erweiterungen zu XML Schema Language benötigt, um allen Anforderungen von MPEG-7 gerecht zu werden.

Kodierung

Dieser Standard ist notwendig um die Beschreibungen in DDL auf geeignete Weise zu kodieren. Dabei soll eine möglichst hohe Kompressionsrate erreicht werden. Die Kodierung soll außerdem

eine hohe Fehlertoleranz bieten und einen wahlfreien Zugriff auf die Beschreibungen in DDL ermöglichen.

4 Arbeitsgruppen

Im folgenden Abschnitt werde ich auf die Arbeiten der Arbeitsgruppen MPEG-7 Audio und MPEG-7 Visual eingehen. Der Hauptaugenmerk liegt dabei auf der Gruppe MPEG-7 Visual, da diese mit ihren Deskriptoren auf Bildern genau die Beschreibungen bieten, die man innerhalb einer Bilddatenbank verwenden kann.

4.1 MPEG-7 Audio

Die AG MPEG-7 Audio entwickelt einerseits generische Audio Deskriptoren, die Beschreibungen der Wellenform oder des Frequenzspektrums bereitstellen. Diese Merkmale können verwendet werden, um ähnliche Stimmen zu finden. Andererseits entwickelt sie auch anspruchsvollere Deskriptoren, wie den Sprachinhalt Deskriptor oder den Timbre Deskriptor.

Mit dem Sprachinhalt Deskriptor soll gesprochenes Audiomaterial beschrieben werden. Der Deskriptor enthält hierbei den Output der neuesten Spracherkennungssoftware, nämlich Wörter und die zugehörigen Phoneme. Die Phoneme werden zusätzlich im Deskriptor gespeichert, damit bei Wörtern, die nicht im Vokabular des Spracherkenners sind, trotzdem die Lautinformationen dieser Wörter vorliegen. Bei einer gesprochenen Suchanfrage, die das gleiche unbekannte Wort enthält, kann so der Audioinhalt gefunden werden, da eine Ähnlichkeit zwischen den Phonemen vorliegt.

Der Timbre Deskriptor ist sinnvoll, da hierbei von der Klangfarbe des Instruments abstrahiert wird. Hierdurch lassen sich ähnliche Melodien finden, obwohl diese von unterschiedlich klingenden Instrumenten eingespielt wurden. Bei einer Suchanfrage durch Summen ist die Klangfarbe durch die Stimme des Suchenden gegeben.

4.2 MPEG-7 Visual

Auch die AG MPEG-7 Audio entwickelt Grundstruktur-Deskriptoren, die z.B. Tools zur Segmentierung und zur Partitionierung und Möglichkeiten zur Interpolation bieten. Der zweite große Teil beschäftigt sich mit den Deskriptoren für visuelle Merkmale. Im folgenden werde ich eine Auswahl von Grundstruktur-Deskriptoren und Deskriptoren für visuelle Merkmale vorstellen.

4.2.1 Grundstruktur Deskriptoren

Grid Layout

Der Grid Layout Deskriptor wird verwendet um ein Bild in gleichmäßige Rechtecke zu teilen. Innerhalb der einzelnen Rechtecke können dann Deskriptoren für visuelle Merkmale den Inhalt beschreiben.

2D/3D Multiple View

Durch diesen Deskriptor wird es möglich ein 3D-Objekt zu beschreiben, indem es durch 2D-Deskriptoren aus verschiedenen Ansichten beschrieben wird. So ist es möglich zu einem 3D-Objekt als Suchanfrage ein ähnliches 3D-Objekt zu finden. Aber auch zu einem 2D-Bild kann man ein 3D-Objekt finden, falls das Bild Ähnlichkeit zu einer Ansicht des 3D-Objekt hat.

Temporal interpolation

Der Temporal Interpolation Deskriptor stellt eine Methode bereit um einen beliebigen Parameter, z.B. die Position eines Objekts in einer Videosequenz, durch eine Interpolation zu beschreiben. In Abbildung 3 ist die Interpolation der x-Position eines Objekt innerhalb einer Videosequenz dargestellt. Beschreibt man die Objektposition in einer 25 Frames langen Videosequenz durch 25 reelle Zahlen, kann man durch eine lineare Interpolation mit 6 Stützstellen schon eine gute Approximation der Trajektorie erreichen. Eine quadratische Interpolation mit 3 Stützstellen würde ebenfalls ein sehr gutes Resultat ergeben, zusätzlich hat man nun eine stetige Trajektorie vorliegen.

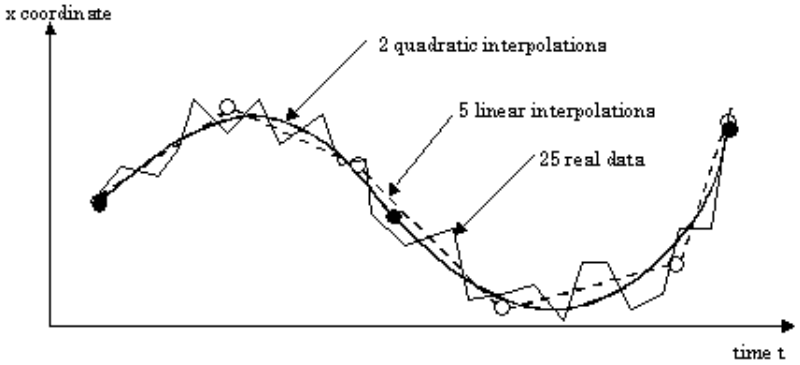

Abbildung 3: Interpolation der x-Position eines Objekts

4.2.2 Deskriptoren für Farbmerkmale

Farbraum

Hierbei ist der Farbraum des Bildes das Merkmal und dieser soll nach Möglichkeit auch in den anderen visuellen Deskriptoren verwendet werden.

Farbquantisierung

Dieser Deskriptor beschreibt eine einheitliche Quantisierung eines Farbraums. Die Anzahl der Farben, auf die quantisiert werden soll, ist frei konfigurierbar. Eine Farbquantisierung ist in Abbildung 4 zu sehen. Die Farbquantisierung ist nützlich, wenn man ein Bild mit Hilfe des Deskriptors für dominante Farben beschreibt.

Dominante Farben

Dieser Deskriptor repräsentiert den prozentualen Anteil jeder quantisierten Farbe in lokalen Bereichen(Objekt oder Bildauschnitt). Er ist bei Bildern einzusetzen, bei denen eine geringe Anzahl von Farben ausreichend für eine Charakterisierung der Farbinformationen ist. Ein Beispiel sind die drei Objekte(Paprikas) in Abbildung 4, die wir durch ihre dominanten Farben (rot, grün, gelb) unterscheiden können.

Skalierbares Farbhistogramm

Dieser Deskriptor ist ein Farbhistogramm im HSV Farbraum, welches durch eine Haar-Transformation kodiert wurde. Wichtig ist hier, dass die binäre Repräsentation des Histogramms in Bin- und Bitanzahl skalierbar ist. Dieser Deskriptor ist geeignet, um ein Bild als Suchanfrage zu bearbeiten und Bilder mit ähnlichem Histogramm zurückzugeben.

Abbildung 4: Beispiel einer Farbquantisierung auf 6 Farben (rechtes Bild)

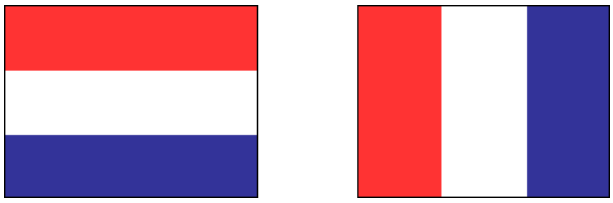

Abbildung 5: Die Bilder der beiden Fahnen sind durch das Farbhistogramm nicht unterscheidbar. Die Struktur der Farbinhalte liefert eine Möglichkeit zur Unterscheidung.

Farb-Layout

Dieser Deskriptor repräsentiert die räumliche Farbverteilung visueller Farbsignale in kompakter Form. Die Kompaktheit erlaubt es mit hoher Effizienz zu Suchen ohne viel Rechenzeit zu beanspruchen. Daher ist auch ein Gebrauch in mobilen Anwendungen, bei denen die Hardwareressoucen im Allgemeinen stark eingeschränkt sind, möglich. Außerdem bietet dieser Deskriptor eine benutzerfreundliche Schnittstelle, um handgemalte Skizzen als Suchanfragen zu bearbeiten.

Farbstruktur

Um nicht nur den Farbinhalt, sondern auch die Struktur der Farbinhalte zu beschreiben, wurde dieser Deskriptor geschaffen. Im Gegensatz zum Farbhistogramm-Deskriptor können zwei Bilder, die zwar das gleiche Histogramm haben, die Farben aber anders zueinander strukturiert sind, voneinander unterschieden werden. Ein Beispiel für dieses Szenario wird in Abbildung 5 dargestellt. Die Fahnen von Frankreich und den Niederlanden unterscheiden sich nur durch die Struktur der Farben. Die Farbanteile der Farben sind jedoch in beiden Bildern gleich.

4.2.3 Deskriptoren für Texturen

Homogene Texturen

Jedes Bild ist im Prinzip ein Mosaik von homogenen Texturen. Wenn man diese Textur-Merkmale mit einer Bildregion verknüpft, können diese Informationen genutzt werden, um ein Bild zu indizieren. Beispiele für homogene Texturen sind Parkplätze, die aus der Vogelperspektive fotografiert wurden, da die Autos eine regelmäßiger Größe und die gleichen Abstände zueinander haben. Ein weiteres Beispiel wäre die Anwendung in einer Satellitenbilder-Datenbank. Dort erhielte man mit einer Anfrage "Suche nach Satellitenbilder von Deutschland mit weniger als 20% Bewölkung" entsprechende Bilder, da auch die Wolken eine homogene Textur darstellen.

Kantenhistogramm

Dieser Deskriptor enthält ein Kantenhistogramm, welches die Häufigkeit von fünf Arten von Kanten enthält. Einmal die 4 gerichteten Kanten (/, \, -, |) und eine ungerichtete Kante, die alle Kanten repräsentiert, die einen ungeraden Verlauf besitzen. Durch diese Beschreibungsform ist ein Finden von Bildern mit ähnlicher Semantik möglich. Als Suchanfrage kann ein Bild oder eine Skizze verwenden werden. Durch Kombination mit dem Farbhistogramm-Deskriptor kann man die Qualität des Suchergbnisses noch verbessern.

4.2.4 Deskriptoren für Formen

Region Shape

Die Form eines im Bild enthaltenen Objektes stellt ein Merkaml dar, das durch den Region Shape Deskriptor repräsentiert wird. Dieser beschreibt die Form durch einzelne (vgl. Abbildung 6 (a,b)) oder durch mehrere (d, e) Regionen. Desweiteren können die Formen auch Löcher enthalten (c).

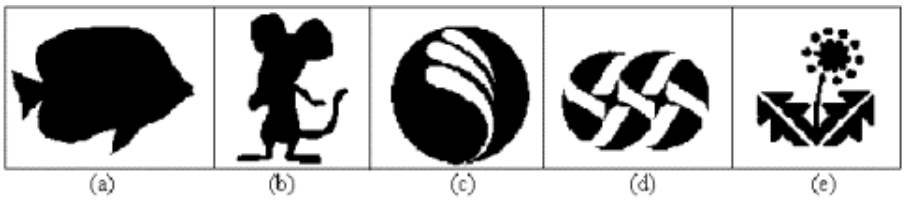

Abbildung 6: Region Shape Beispiele

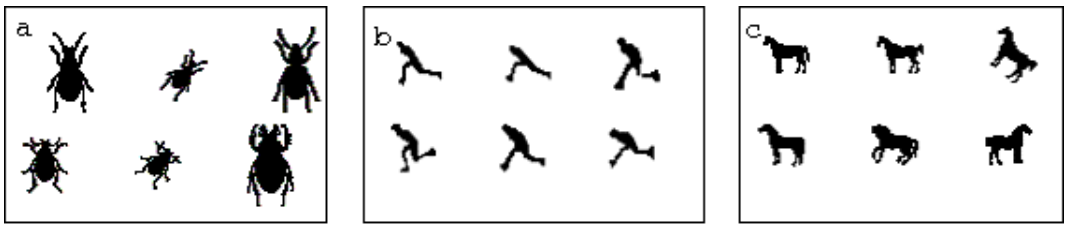

Abbildung 7: (a) es wird eine Ähnlichkeit gefunden, obwohl die Formen sich unterscheiden (ähnlich menschlicher Wahrnehmung), (b) Robustheit bei nicht-starren Bewegungen, (c) Robustheit bei teilweiser Verdeckung

Contour Shape

Eine komplexere Möglichkeit, die Formmerkmale eines Bildes zu beschreiben, bietet der Contour Shape Deskriptor. Dieser basiert auf dem "Curvature Space Scale". Dieses Verfahren hat den Vorteil, dass es charakteristische Formverläufe aufnimmt und ähnlich der menschlichen Wahrnehmung funktioniert. Außerdem ist das Verfahren sehr robust gegenüber nicht-starren Bewegungen, teilweiser Verdeckung und einer Änderung der Perspektive. Abbildung 7 zeigt ein Suchergebnis aus der MPEG-7 Shape Database.

4.2.5 Gesichtserkennung

Der Gesichtserkennungs-Deskriptor kann zum Finden von Gesichtern verwendet werden. Hierzu wird ein Bild normalisiert, indem ein Raster mit 56 x 46 Quadraten so auf das Bild gelegt wird, dass die Schnittpunkte der 16. und 31. Spalte mit der 24. Reihe in den Augenmittelpunkten liegen. Die Normalisierung eines Gesichts wird in Abbildung gezeigt. Dieses Raster bildet durch die Helligkeitswerte in den einzelnen Quadraten einen Vektor, den man nun auf die Gesichtsbasisvektoren projezieren muss. Die Menge der Werte der Projektionen bilden die Merkmalsmenge des Gesichts.

Abbildung 8: Gesichtserkennung: Das Bild muss zunächst durch eine definierte Methode normalisiert werden. Dabei wird ein Raster so auf das Bild gelegt, dass die Schnittpunkte der roten Geraden in den Augenmittelpunkten liegen.

5 Anwendungen

5.1 PicSOM Image Query

Diese Anwendung verwendet intern das MPEG-7 Format. Weitere Infos unter http://www.cis.hut.fi/picsom

5.2 IBM MPEG-7 Annotation Tool

Dies ist ein Assistent zur manuellen Annotation von Videomaterial in MPEG-7 Format. Die Szenen des Videos werden automatisch beim Einladen der Datei segmentiert, und können dann einzeln manuell bearbeitet werden.

5.3 Ricoh MPEG-7 MovieTool

Dieses Tool ist dem IBM Annotation Tool sehr ähnlich. Auch hier werden die einzelnen Szenen automatisch detektiert. Dies geschieht hier allerdings während man das Video anschaut, also in Echtzeit. Das Programm bietet einen MPEG-7 Editor, der die Syntax der MPEG-7 Beschreibungen auf Korrektheit überprüft.

5.4 Canon MPEG-7 Speech Recognition engine

Diese Anwendung gibt als Ausgabe eine MPEG-7 Beschreibung des gesprochenen Audiomaterials. Im Moment werden allerdings nur die Phoneme klassifiziert und die Erkennungsrate liegt bei 60%.

Literatur

[1] MPEG-7 Introduction:
http://ipsi.fraunhofer.de/delite/Projects/MPEG7/Documents/W4325%20M7%20Intro.htm

[2] MPEG-7 Overview:
http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

[3] MPEG-7 Industry:
http://www.mpeg-industry.com/