*Keyframe Extraktion für Video-Annotation und Video-Zusammenfassung**

Damian Borth · Adrian Ulges
Christian Schulze · Thomas M. Breuel

Einleitung

Möchte man Videos im Internet suchen, benutzt man oft die vorhandenen Suchmechanismen bekannter Videoportale. YouTube als Marktführer, bietet auf seiner Webseite eine Schlagwortsuche an, welche mittels manuell erzeugter Meta-Information und hinzugefügter Tags arbeitet. Leider sind diese Meta-Information und Tags in ihrer Fähigkeit limitiert, den Inhalt eines Videos vollständig zu beschreiben bzw. die verwendeten Tags sind subjektiv und können in ihrer Semantik irreführend sein. Aus diesen Grund können inhaltsbasierende Methoden wie „Content-based Video Retrieval" (CBVR) die Videosuche entscheident verbessern. Eine Schlüsselkomponente des CBVR ist die Extraktion von sogenannten „Keyframes". Dies sind Einzelbilder, welche den Inhalt eines Videos bestmöglich wiedergeben und sich damit für Bildanalysen mittels bekannten Bildverarbeitungsalgorithmen besonders eignen. Aus diesen Keyframes werden im Analyseprozess visuelle Merkmale extrahiert, die für eine weitere Verwendung in einer Datenbank indexiert werden können. Unser Beitrag zeigt, dass die Qualität der extrahierten Keyframes die Gesamtleistung eines CBVR Systems signifikant beeinflussen kann. Die extrahierten Keyframes können auch dazu benutzt werden, den Inhalt eines Videos zusammenzufassen und dadurch den Benutzer zu helfen einen schnellen Überblick über die Relevanz seines Suchergebnisses zu erhalten. Dabei können die extrahierten Keyframes entweder als Mosaik dargestellt werden [1] oder für die Navigation auf eine temporär-visuellen Ebene gemappt werden [7].

Unser Ansatz

Video ist ein temporäre strukturiertes Medium. Deswegen ist der erste Schritt im Vorverarbeitungsprozess eines CBVR Systems die temporäre Segmentierung eines Videoclips in seine zugrundelegenden Kameraeinstellungen (Shots). Basierend auf verschiedenen Ansätzen zur Extrahierung von Keyframes können Repräsentanten für die erkannten Shots ermittelt werden. Ein einfacher Ansatz ist das Definieren des ersten, mittleren oder letzten Bildes eines Shots als Keyframe [6]. Dies kann jedoch bei komplexeren Shots zu Informationsverlust führen. Um dieses Problem zu lösen, haben Hammoud und Mohr [2] den Einsatz von unüberwachten Lernenmethoden vorgeschlagen, welche sich der Komplexität des Shots anpassen. Unser Ansatz beschreibt eine Kombination aus einer „Shot-Segmentierung" auf dem Videoclip und eines „Intra-Shot Clusterings" auf den Bildmaterial eines jeden Shots, um eine adäquate Anzahl an Keyframes in Bezug zur visuellen Komplexität des Videos zu extrahieren.

*Englische Originalversion unter: http://pubs.iupr.org

DOI 10.1007/s00287-008-0264-y

Damian Borth · Adrian Ulges · Thomas M. Breuel
Technische Universität Kaiserslautern,
Image Understanding & Pattern Recognition,
67663 Kaiserslautern
E-Mail: d_borth@informatik.uni-kl.de
www.iupr.org

Christian Schulze · Thomas M. Breuel
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI),
Image Understanding & Pattern Recognition,
67663 Kaiserslautern

Zusammenfassung

Zur Zeit erleben Videoclips, welche auf online Videoportalen wie YouTube zur Verfügung gestellt werden, immer mehr an Popularität. Wir schlagen einen Ansatz vor, der basierend auf unüberwachtem Lernen, Keyframes für Video-Retrieval und Video-Zusammenfassungen extrahiert. Unser Ansatz nutzt Methoden der „Shot-Segmentierung" um ein Video temporär zu segmentieren und einen „k-Means" Algorithmus um Repräsentanten für jeden Shot zu bestimmen. Zusätzlich führen wir ein „Meta-Clustering" auf den extrahierten Keyframes aus um kompakte Videozusammenfassungen zu erhalten. Um unsere Methoden zu testen haben wir diese auf einer Datenbank von YouTube Videos angewendet. Wir erhielten Ergebnisse, welche (1) eine Verbesserung des Retrievals und (2) kompakte Video-Zusammenfassungen zeigen.

Shot-Segmentierung. Neben den vielen bekannten Methoden der temporären Segmentierung [3] benutzen wir Differenzen des MPEG-7 „Color Layout Descriptors" (CLD), welcher als visuelles Merkmal aus zwei aufeinanderfolgenden Einzelbildern extrahiert wurde. Das Erkennen von Shotgrenzen auf den Differenzfolgen wird mittels einer adaptiven Schwellentechnik [4] erreicht. Diese Methode funktioniert sehr gut bei harten Schnitten – jedoch verliert sie an Stärke bei längeren Blenden und Übergängen.

Intra-Shot Clustering. Wir modellieren jeden Shot als Menge von Gaussverteilungen, welche wir mittels des „k-Means" Algorithmuses [5] ermitteln. Das Clustering wird auf CLD Merkmalen der Einzelbilder eines jeden Shots durchgeführt. Danach wird das Einzelbild, welches am Nächsten zum Mittelpunkt eines jeden Clusters liegt, als Repräsentant des Clusters definiert. Um die Anzahl der Cluster für k-Means zu ermitteln, benutzen wir das bayessche Informationskriterium (BIC) [8] als Validierungsmaß für die Güte des Clusterings. Dies kann, bedingt durch die Komplexität des Shots, zu mehreren Clusters, sprich Keyframes, pro Shot führen. Abbildung 1 zeigt als Beispiel das Clustering eines komplexeren Shots (aus Visualisierunggründen auf 2d reduziert). Visuell ähnliche Bilder sind markiert und deren Repräsentanten am linken und rechten Rand angezeigt.

Keyframe „Meta-Clustering". Ein Nachteil des zuvor beschriebenen zweigeteilten Prozesses ist die redundante Auswahl von Keyframes bedingt durch fehlende Inter-Shot Information. Dies kann in Dialogszenen besonders oft passieren wie auch in Musikvideos, bei den der Künstler innerhalb des Videos mehrfach gezeigt wird. Eine solche redundante Auswahl von Keyframes ist bei Video-Zusammenfassungen nicht gewünscht und wird bei unserem System mittels

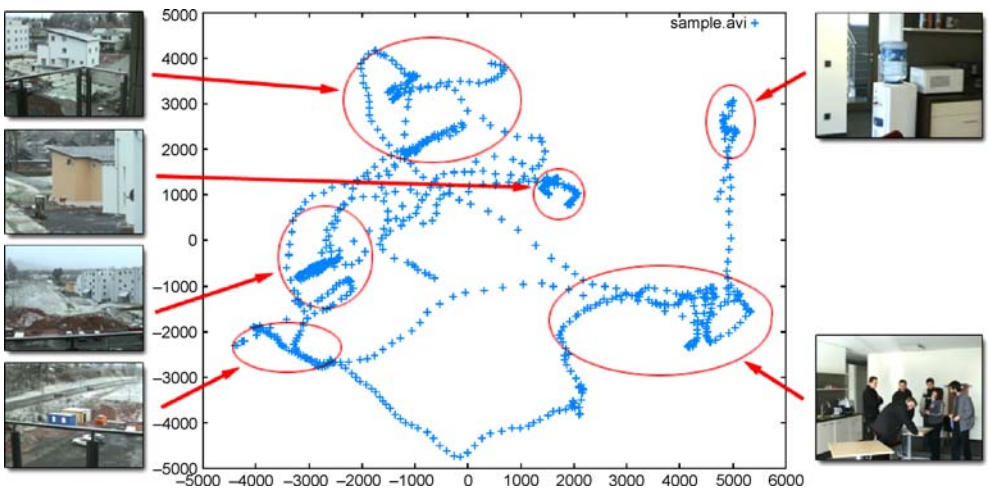

Abb. 1 Clustering eines komplexeren Shots (Mitte). Aus Visualisierungsgründen mittels PCA auf 2d reduziert. Das Clustering führt zu den auf der linken und rechten Seite dargestellten Keyframes

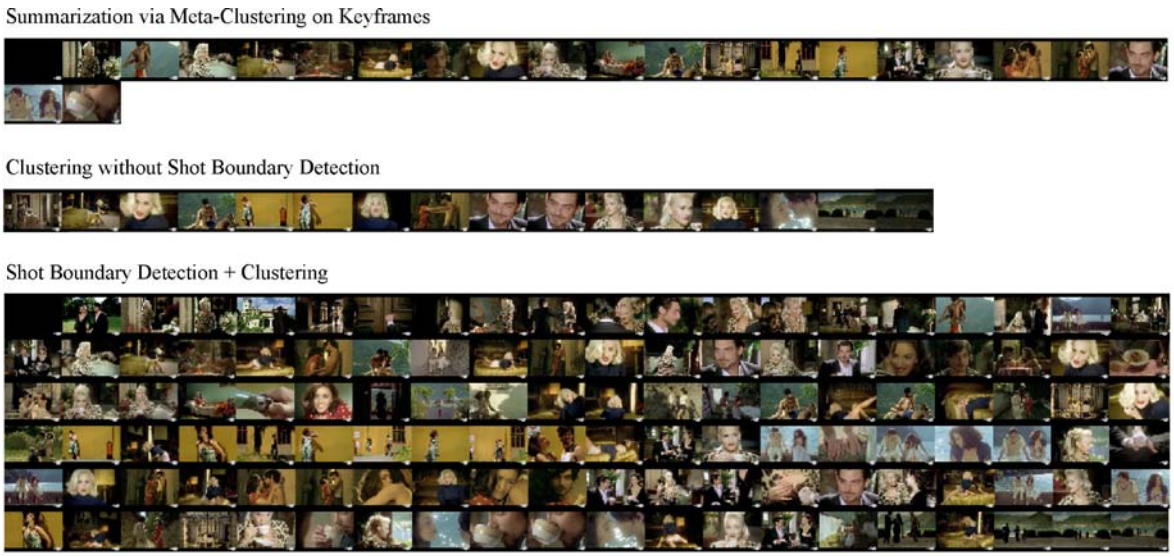

Abb. 2 Ergebnisse einer Videozusammenfassung für ein Musikvideo. Die erste Zeile zeigt das Ergebnis des „Meta-Clustering". Die zweite Zeile zeigt ein reines Clustering auf dem Video. Die dritte Zeile zeigt die extrahierten Keyframes an

eines „Meta-Clusterings" auf den zuvor extrahierten Keyframes kompensiert. Ähnlich wie [1] verwenden wir die Idee einer erneuten Keyframegruppierung um Video-Zusammenfassungen zu erzeugen. Aber im Kontrast zu dem Erzeugen einer hierarchischen Struktur mit Shots und Szenen, benutzen wir die bereits vorhandenen Keyframes und clustern diese um Redundanz zu vermeiden.

Experimente & Ergebnisse

Annotation

Die beschriebene Keyframe-Extrahierungsmethode wurde als Vorverarbeitungsschritt innerhalb eines Experimentes zur automatischen Annotation einer Videodatenbank mit über 2200 Videos benutzt. Diese Videodatenbank besteht aus Videos heruntergeladen von YouTube und hat eine Abspiellänge von 194 Stunden. Bei dem Experiment lernte ein „Video Tagger"[1] Konzepte wie *Katze*, *Segeln* und *Wüste* auf einer Trainingsmenge von durch YouTube Benutzern zuvor annotierten Videos und erreichte eine „Mean Average Precision" (MAP) von 34.2% auf unbekannten Videos. Es wurden verschiedene Kombinationen von visuellen Merkmalen wie Farbhistogrammen, Tamura Merkmale und Visuelle Wörter bei diesem Experiment aus dem Keyframes extrahiert. Zusätzlich wurden auch Bewegungsmerkmale auf Shot Ebene extrahiert. Um den Einfluss verschiedener Keyframe-Extrahierungsmethoden für eine erfolgreiche Annotation zu messen wurde das Experiment mit zwei weiteren Extrahierungsmethoden durchgeführt: `[Erstes-Bild]`, hier nehmen wir das erste Bild als Repräsentanten für den Shot und `[Reg-Sampling]`, bei dem wir regulär in 7 Sekunden Intervall Keyframes aus jedem Videoclip extrahieren. Abhängig vom der gewählten Kombination von visuellen Merkmalen zeigt unserer Keyframe-Extrahierungsmethode eine Verbesserung von 2–9% verglichen mit `[Erstes-Bild]` und eine ähnliche Leistung wie `[Reg-Sampling]`, jedoch mit weniger Keyframes.

[1] im Detail beschrieben in [9]

Video-Zusammenfassung

Das „Meta-Clustering" wurde auf verschiedenen Musikvideos getestet, da sich diese oft durch eine hohe Anzahl an Schnitten und redundanten Shots auszeichnen. Durch das verwendete „Meta-Clustering" und die dadurch erzielte Gruppierung von mehrfach vorkommenden Keyframes erhielten wir kompakte Videozusammenfassungen. Ein pures Clustering auf dem gesamten Videoclip, welches als zusätzliches Experiment durchgeführt wurde, führte zu der ungefähr gleichen Anzahl an Keyframes, hat aber inhaltlich oft nicht das gesamte Video wiedergeben können. Abbildung 2 zeigt ein Musikvideo als Beispiel, bei dem redundante Keyframes erfolgreich zusammengefasst wurden. Die

erste Zeile des Beispieles zeigt das Ergebnis des „Meta-Clustering“, welches auf den zuvor extrahierten Keyframes durchgeführt wurden. Diese sind in der dritten Zeile dargestellt werden. Die zweite Zeile zeigt extrahierte Keyframes, welche durch ein pures Clustering auf dem Gesamtvideo ermittelt wurde. Wir haben vergleichbare Ergebnisse auch für anderes Videos erhalten.

Literatur

1. Bailer W, Mayer H, Neuschmied H, Haas W, Lux M, Klieber W (2003) Content-based video retrieval and summarization using MPEG-7. In: Santini S, Schettini R (Hrsg) Internet Imaging V. Proc SPIE 5304:1–12
2. Hammoud R, Mohr R (2000) A Probabilistic Framework of Selecting Effective Key-Frames for Video Browsing and Indexing. In: Intern. Worksh. on Real-Time Img. Seq. Anal., S 79–88
3. Koprinska I, Carrato S (2001) Temporal Video Segmentation: A Survey. Signal Process Image Commun 16:477–500
4. Lienhart R (2001) Reliable Transition Detection in Videos: A Survey and Practitioner's Guide. Int J Image Graph 1(3):469–486
5. McQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, S 281–297
6. O'Connor B (1991) Selecting Key Frames of Moving Image Documents: A Digital Environment for Analysis and Navigation. Microcomput Inf Manage 8(2): 119–33
7. Rautiainen M, Ojala T, Seppanen T (2004) Cluster-temporal browsing of large news video databases. IEEE Int Conf Multimed Expo 2:751–754
8. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
9. Ulges A, Schulze C, Keysers D, Breuel TM (2007) Content-Based Video Tagging for Online Video Portals. In: MUSCLE/Image-CLEF Workshop