OSOTIS – Kollaborative inhaltsbasierte Video-Suche

Harald Sack, Jörg Waitelonis

Friedrich-Schiller-Universität Jena
D-07743 Jena
{sack, joerg}@minet.uni-jena.de

Abstract: Die Video-Suchmaschine OSOTIS ermöglicht eine automatische inhaltsbezogene Annotation von Videodaten und dadurch eine zielgenaue Suche auch innerhalb einzelner Videoaufzeichnungen. Neben objektiv gewonnenen zeitabhängigen Deskriptoren, die über eine automatische Synchronisation von ggf. zusätzlich vorhandenem textbasiertem Material mit den vorliegenden Videodaten gewonnen werden, können kollaborativ zusätzlich eigene, zeitbezogene Schlagwörter (Tags) und Kommentare innerhalb eines Videos vergeben werden (sequentielles Tagging), die zur Implementierung einer verbesserten und personalisierten Suche dienen.

1 Einleitung

Die Informationsfülle des World Wide Webs (WWW) ist gewaltig. Milliarden von Dokumenten in hunderten von Sprachen machen es unmöglich, sich ohne Hilfsmittel darin zu orientieren. Suchmaschinen wie Google[1] verfolgen das Ziel, den erreichbaren Teil des WWWs, möglichst vollständig zu indizieren und so durchsuchbar zu machen. Noch immer stellen Textdokumente den größten Anteil des WWWs dar, aber immer mehr Multimedia-Dokumente in Form von Bildern, Grafiken oder Video-Clips kommen täglich hinzu. Google allein verwaltet derzeit in seinem Suchindex mehr als 1,2 Milliarden Bilder und mehrere Millionen Videos (Stand: 05/2007).

Insbesondere der Anteil an Videodaten im WWW steigt auf Grund vielfältiger Content Management Systeme zur Produktion, Nachbearbeitung und Bereitstellung, sowie der stetig wachsenden zur Verfügung stehenden Bandbreite. Spezialisierte Portale und Video-Suchmaschinen wie etwa YouTube[2] oder Google Video[3] erleichtern das Auffinden von Videodaten im WWW. Gegenüber traditionellen Suchmaschinen, d. h. Suchmaschinen für textbasierte Dokumente, unterscheiden sich Video-Suchmaschinen typischerweise in der Art der Indexerstellung. Traditionelle Suchmaschinen wenden Methoden des Information Retrieval auf Textdokumente an, um aus diesen aussagekräftige Deskriptoren zur Beschreibung und Verschlagwortung des untersuchten Dokuments zu gewinnen. Diese vollautomatische Suchindexgenerierung ist im Falle von multimedialen Daten in der Regel

[1] Google, http://www.google.com/
[2] YouTube, http://www.youtube.com/
[3] Google Video, http://video.google.com/

schwierig oder erst gar nicht möglich. Mit klassischen Methoden des Information Retrieval angewandt auf multimediale Daten ist es lediglich möglich, charakteristische Eigenschaften wie z. B. dominante Farben, Farb- und Helligkeitsverteilungen in Einzelbildern oder die Bewegungen der Kamera innerhalb einer Bildfolge zu bestimmen bzw. einzelne Objekte zu identifizieren oder zu verfolgen. Zwischen diesen charakteristischen Eigenschaften und dem tatsächlichen Inhalt der multimedialen Daten und dessen Bedeutung besteht eine semantische Lücke [Sm00]. Schlussfolgerungen aus den charakteristischen Eigenschaften auf deren inhaltliche Bedeutung sind heute nur in geringem Maße möglich. Ebenso ist eine automatische Extraktion inhaltsbezogener Deskriptoren, die den semantischen Inhalt einer Videodatei auf einer abstrakteren Ebene beschreiben, aus den Videodaten allein nicht zufriedenstellend möglich.

Die inhaltliche Beschreibung multimedialer Daten und insbesondere von Videodaten erfolgt über eine Annotation mit zusätzlichen Metadaten, die entweder vom Autor der Daten selbst, von ausgewiesenen Experten oder aber auch von allen Nutzern gemeinsam erfolgen kann. Letztere sind auch verantwortlich für den Erfolg von Web-2.0-Video-Suchmaschinen wie YouTube, da diese dem Nutzer eine einfache Annotation der Videos über das so genannte Tagging ermöglichen, d. h. die Nutzer vergeben eigene, frei gewählte Schlüsselwörter (Tags), die den Inhalt der Videodaten beschreiben.

Betrachtet man speziell den Anteil an Lehr- und Lernmaterialien in Video-Suchmaschinen, ist dieser heute sehr gering. Dies hat verschiedene Gründe: Einerseits liegen Lehr- und Lernmaterialien oft auf spezialisierten Portalen oder Lernplattformen vor, die entweder aus den bereits oben genannten Gründen bzw. auf Grund eines dezidierten Rechtemanagements nicht von Video-Suchmaschinen indiziert werden können. Andererseits liegt ein weiteres Problem in der Natur der Videomaterialien selbst begründet: Die Videoaufnahme einer Lehrveranstaltung hat in der Regel eine Länge zwischen 45 und 90 Minuten. Dabei werden in einer Lehrveranstaltung oft unterschiedliche Themen behandelt. Einzelne Themen nehmen in der gesamten Lehrveranstaltung oft nur wenige Minuten in Anspruch und sind nur schwer darin wiederzufinden. Zwar können durch Autor oder Nutzer Tags bereitgestellt werden, die alle in der Vorlesung angesprochenen Themen beschreiben, doch ist deren zeitliche Zuordnung innerhalb des zeitgebundenen Mediums Video ebenso wie eine direkte zeitliche Adressierung bei der Wiedergabe der Suchergebnisse noch nicht realisiert.

Im vorliegenden Beitrag beschreiben wir die Video-Suchmaschine OSOTIS[4], die eine zeitabhängige, sequentielle Indizierung von Videodaten und damit eine direkte Suche auch innerhalb dieser Videodaten ermöglicht. Insbesondere dient OSOTIS dabei der Archivierung und der Annotation von videobasierten Lehr- und Lernmaterialien, wie z. B. Vorlesungsaufzeichnungen. OSOTIS kombiniert zwei unterschiedliche Ansätze: Zum einen werden Vorlesungsaufzeichnungen, zu denen eine Desktopaufzeichnung des Dozenten und zusätzliche Daten wie z. B. eine Präsentation, ein Handout oder eine Vorlesungsmitschrift vorliegen, automatisch mit dem Inhalt dieser Zusatzinformationen synchronisiert und annotiert. Zum anderen gestattet OSOTIS jedem Benutzer die Vergabe von zeitabhängigen Tags, d. h. eine bestimmte Stelle des Videos kann während des Abspielens von den Nutzern mit eigenen Tags oder ganzen Kommentaren annotiert werden, die dann wieder

[4] OSOTIS, http://www.osotis.com/

gezielt abgerufen werden können. Eigene Tags ermöglichen dem Benutzer eine personalisierte Suchfunktion und mit Hilfe der gemeinsamen Tags aller übrigen Benutzer wird die herkömmliche Suche ergänzt. OSOTIS bietet dem Benutzer die Möglichkeit, aus einem stetig wachsenden Datenbestand an Vorlesungs- und Lehrvideos, zielgerichtet und nach persönlichen Vorgaben, eigene Vorlesungen aus einzelnen Videosequenzen entsprechend seinen persönlichen Bedürfnissen zusammenzustellen.

Nachfolgend soll die Arbeitsweise von OSOTIS detaillierter beschrieben werden: Kapitel 2 untersucht Eigenschaften und Defizite aktueller Video-Suchmaschinen. Kapitel 3 zeigt die Möglichkeiten einer automatischen Annotation von Video-Daten, während Kapitel 4 näher auf die kollaborative Annotation zeitabhängiger Daten eingeht. Kapitel 5 gibt einen Einblick in die Arbeitsweise der Video-Suchmaschine OSOTIS und Kapitel 6 beschließt die Arbeit mit einem kurzen Ausblick auf deren Weiterentwicklung.

2 Aktuelle Video-Suchsysteme

Video-Suchsysteme können auf unterschiedliche Art zu dem in ihnen repräsentierten Datenbestand gelangen: Crawler-basierte Systeme durchsuchen in der Art traditioneller Suchmaschinen das WWW aktiv nach Videodaten und verwenden zum Aufbau ihres Suchindexes neben den aufgefundenen Videodaten ebenfalls verfügbare Kontextinformation (z. B. Hyperlink-Kontext bei Google Video). Upload-basierte Systeme ermöglichen registrierten Nutzern als Publikationsplattform das Einstellen eigener Videodaten (z. B. YouTube). Daneben existieren redaktionell gepflegte Systeme, die es lediglich einem ausgewählter Kreis von Nutzern ermöglichen, eigenes Videomaterial einzustellen (z. B. Fernsehsender, Nachrichtenredaktionen und digitale Bibliotheken[5] an Universitäten und anderen Bildungseinrichtungen).

Analog zu traditionellen Suchmaschinen können auch im Falle von Video-Suchmaschinen indexbasierte Suchmaschinen und Suchkataloge unterschieden werden. Indexbasierte Suchmaschinen liefern auf die Eingabe eines oder mehrerer Suchbegriffe eine nach internen Relevanzkriterien hin sortierte Ergebnisliste. Viele redaktionell gepflegte Systeme dagegen arbeiten nach dem Prinzip des Suchkatalogs, d. h. sie erlauben lediglich das Blättern und Navigieren in vordefinierten Kategorien. überschreitet das angebotene Videomaterial eine bestimmte Dauer, ist eine inhaltsbasierte Recherchemöglichkeit unverzichtbar.

Inhaltsbasierte Suche nach und in Videodaten erfolgt nach unterschiedlichen Kriterien. Man unterscheidet hier die Suche über Kategorien, Schlüsselwörter, Schlagworte/Tags, eine semantische Suche, Suche nach analytischen Bildeigenschaften oder die Suche nach dem gesprochenen Wort. Aktuelle Suchmaschinen stellen kategorien- und schlüsselwortbasierte Suche sowie die Suche nach Tags bereit. Des weiteren kann nach der Suchgranularität unterschieden werden. Dies betrifft Sammlungen (Kollektionen) von Videos, ein einzelnes Video, ein Videosegment, eine Szene (Group of Pictures), den Teilbereich einer Szene (Objekt-Verfolgung), ein Einzelbild oder den Teilbereich eines Einzelbildes. Die aktuellen Video-Suchdienste wie Google-Video und YouTube sind lediglich in der

[5] z. B. Digitale Bibliothek Thüringen, http://www.db-thueringen.de

Lage, nach einzelnen Videos als Ganzem zu suchen. Einen Ansatz mit feinerer Granularität verfolgen die Systeme TIMMS[6], Slidestar[7] und OSOTIS. Mit diesen Systemen ist es möglich, auch den Inhalt einzelner Videos zu durchsuchen. Die Unterschiede zwischen den Systemen liegen in der Medienaufbereitung und Metadatengewinnung. Während bei TIMMS Videodaten manuell mit großem Aufwand segmentiert und annotiert werden, verwendet Slidestar das proprietäre Lecturnity[8] Format, um eine automatische Indizierung der Videodaten zu realisieren. Dazu müssen Metadaten wie Folientext und Autorenannotationen bereits während der Produktion in das Lecturnity Format eingebettet werden, um von Slidestar zur inhaltsbasierten Suche genutzt werden zu können. Dagegen ist es mit OSOTIS möglich, beliebige Videoformate mit vorhandenem textuellen Präsentationsmaterial (z. B. im PDF[9] oder PPT[10] Format) vollautomatisch zu resynchronisieren, um positionsabhängige Metadaten zu generieren, die die Grundlage für die Indizierung bilden [SW06a].

Aus Effizienzgründen erstellen Suchmaschinen einen Suchindex, der einen schnellen Zugriff auf die Suchergebnisse mit Hilfe von Deskriptoren gestattet, die direkt aus den zu durchsuchenden Daten bzw. aus zusätzlichen Metadaten (Annotationen) gewonnen werden. Deskriptoren sind zum einen analytische/syntaktische Merkmale (z. B. Farbe, Form, Objekte), semantische Eigenschaften (z. B. Beziehungen zwischen Objekten) oder auch Zusatzinformationen. Der Grad an Automatisierbarkeit bei der Erzeugung der Deskriptoren fällt in der genannten Reihenfolge ab. Deskriptoren können sich dabei auf einzelne Teile der Videodaten (z. B. Videosegmente, Einzelbilder, Bereiche) beziehen.

Zur Ermittlung geeigneter Deskriptoren für den speziellen Fall der Suche in Aufzeichnungen von Lehrveranstaltungen stehen inhaltliche, semantische Gesichtspunkte im Vordergrund, also z. B. welches Thema wird zu welchem Zeitpunkt oder in welchem Videosegment behandelt. Allerdings enthält der Videodatenstrom einer Lehrveranstaltungsaufzeichnung keine geeigneten charakteristischen Merkmalsausprägungen über den Zeitverlauf hinweg. Jedes einzelne Videosegment ähnelt jedem anderen visuell so stark – in den meisten Fällen ist ausschließlich ein Vortragender zu sehen – dass bei alleiniger Betrachtung eines einzelnen Videosegments oft nicht festzustellen ist, zu welchem Zeitpunkt der Aufzeichnung dieses gehört. Objektidentifikation, Objektverfolgung und eine Segmentierung entsprechend der Schnittfolge eines Videos sind in diesem Falle ebenfalls nicht sinnvoll, da nicht auf den semantischen Inhalt der Vorlesung geschlossen werden kann, höchstens auf eine Person, die sich z. B. nach links oder rechts bewegt. Merkmalausprägungen von besserer Separierungsfähigkeit können aus den zugehörigen Audiodaten gewonnen werden. Eine Segmentierung kann in diesem Fall z. B. bzgl. der Sprechpausen erfolgen. Die einzelnen Audio-Segmente werden hierzu einer automatischen Sprachanalyse unterzogen, deren Ergebnis die gewünschten Merkmale hervorbringt (vgl. Kap. 3).

Systeme, die Aufzeichnungen von Lehrveranstaltungen verwalten, müssen in der Lage sein, auch den Inhalt einzelner Videos zu durchsuchen. Lehrveranstaltungen stellen beson-

---

[6] Tübinger Internet Multimedia Server, http://timms.uni-tuebingen.de/

[7] Slidestar IMC AG, http://www.im-c.de/Produkte/170/4641.html. Eine Beispielanwendung ist das eLecture Portal der Universität Freiburg: http://electures.informatik.uni-freiburg.de/catalog/courses.do

[8] Lecturnity IMC AG, http://www.lecturnity.de/

[9] Adobe - Portable Document Format, nahezu alle textuellen Formate lassen sich in das PDF umwandeln.

[10] Microsoft PowerPoint

dere Ansprüche an ein Retrievalsystem. Traditionelles Multimedia Retrieval, das versucht charakteristische, statistische Merkmale zu indizieren, ist in diesem Falle nicht geeignet.

3 Automatische Annotation von Video-Daten

Lehrveranstaltungsaufzeichnungen bestehen heute oft aus synchronisierten Multimediapräsentationen, die eine Videoaufzeichnung des Dozenten, eine Aufzeichnung der Präsentation des Dozenten und einen Audiodatenstrom beinhalten (siehe Abb. 1). Diese können z. B. mit Hilfe der Standards „Synchronous Multimedia Integration Language“[11] (SMIL) oder „MPEG-4 XML-A/O“ [ISO05], aber auch über andere, proprietäre Technologien[12] kodiert werden. Eine synchronisierte Multimediapräsentation enthält bedeutend mehr Informationen als die Videoaufzeichnung des Vortragenden allein. Diese zusätzliche Information wird von OSOTIS genutzt, um eine Vorlesungsaufzeichnung über automatisch generierte Annotationen in eine durchsuchbare Form zu bringen.

Abbildung 1: Synchronisierte Multimediapräsentation bestehend aus Dozentenvideo, Desktopaufzeichnung und interaktivem Inhaltsverzeichnis (links) in Verbindung mit kollaborativem Tagging (rechts) als Ergebnis einer OSOTIS Suchoperation.

Mit einer Aufzeichnung der Präsentation des Dozenten (Desktopaufzeichnung) geht die Verwendung von textuellem Präsentationsmaterial[13] einher. Die aus dem synchronisierten Präsentationsmaterial gewonnene Annotation enthält alle wichtigen Informationen, die über den Inhalt des Videos in Erfahrung zu bringen sind. Die Annotation schließt neben textbasierten, inhaltlichen Zusammenfassungen, Stichpunkten und Beispielen auch Vorschaubilder und andere Multimediainhalte mit ein.

[11] SMIL – Synchronized Multimedia, http://www.w3.org/AudioVideo/
[12] z. B. Lecturnity IMC AG, http://www.lecturnity.de/
[13] z. B. Adobe PDF, Microsoft PowerPoint, o.a.

Der Prozess der Annotation erfolgt entweder bereits online während der Produktion (wie in [ONH04] gefordert) oder auch offline in einem Nachverarbeitungsschritt. Soll eine automatische online-Annotation erfolgen, ist das Führen einer speziellen Log-Datei auf dem Präsentationsrechner des Dozenten erforderlich, in der Ereignisse wie z. B. Folienwechsel aufgezeichnet werden. Aus dieser Log-Datei lässt sich leicht eine zeitliche Synchronisation zwischen Videoaufzeichnung und textuellem Präsentationsmaterial gewinnen. Die Zeitpunkte der jeweiligen Folienwechsel segmentieren die Videoaufzeichnung und die textuellen Inhalte einer Folie werden dem Videosegment als Deskriptor zugeordnet. Textauszeichnungen wie z. B. Schriftschnitt sowie Textposition innerhalb einer Folie (z. B. Kapitelüberschrift) werden dabei zur Relevanzgewichtung der Deskriptoren herangezogen.

Oft ist das Führen einer Log-Datei auf dem Präsentationsrechner nicht möglich oder auch nicht erwünscht. In diesem Fall oder auch für den Fall der Aufbereitung von bereits archiviertem Videomaterial, muss ein analytisches (Retrieval-)Verfahren zur Synchronisation von Videoaufzeichnung und textbasiertem Material verwendet werden. Dies erfolgt bei OSOTIS über Schrifterkennung (Intelligent Character Recognition, ICR) und Bildvergleichanalyse (vgl. [SW06a] für eine ausführlichere Beschreibung der technischen Details). Wird ein ICR-Verfahren allein auf die Präsentationsaufzeichnung angewendet, liefert diese auf Grund oft unzureichender Videoqualität nur eine fehlerhafte Analyse der darin enthaltenen Information [NWP03, KHE05]. Dennoch ist die Qualität dieser Information ausreichend, um eine Synchronisation von Videoaufzeichnung und textuellem Präsentationsmaterial zu gewährleisten. Sollten dabei auf einer Folie keine Textinhalte sondern lediglich Illustrationen und Grafiken enthalten sein, löst ein einfacher analytischer Bildvergleich[14] des Präsentationsmaterials mit der Präsentationsaufzeichnung diese Aufgabe.

Neben dieser bereits realisierten Synchronisation steht derzeit die direkte Synchronisation von Vorlesungsaufzeichnungen mit zusätzlich vorhandenem textuellem Material im Vordergrund der Entwicklung (vgl. [Re07]). Diese Synchronisation basiert auf einer automatischen Spracherkennung (ASR) der aufgezeichneten Audiodaten [CH03, YOA03]. Das Verfahren unterscheidet sprecherabhängige und sprecherunabhängige Spracherkennung. Sprecherabhängige ASR (z. B. Dragon Naturally Speaking[15]) sieht eine Trainingsphase des Systems auf einen bestimmten Sprecher vor. Da eine derartige Trainingsphase des Systems sehr aufwändig ist und mit wachsendem Datenbestand nicht skaliert, liegt der Schwerpunkt der Entwicklung derzeit in der Weiterentwicklung einer sprecherunabhängigen Spracherkennung (z. B. SPHINX [Hu93]). Aktuelle Systeme zur Spracherkennung erreichen eine Fehlerrate (word error rate) von etwa 10 % für englischsprachige[16] und etwa 20 % für deutschsprachige[17] Texte. Zur Verbesserung der Erkennungsrate wird daher ein vorab definiertes, reduziertes Vokabular (Korpus) aus Fachbegriffen zu jeder Vorlesung bereitgestellt, die im Audiodatenstrom zeitlich lokalisiert werden (Term Spotting) [KY96]. Dieses Korpus kann etwa aus dem textuellen Präsentationsmaterial oder aus einer Sammlung von dem Wissensgebiet zugehöriger Fachbegriffe (Lexikon, Ontologien) generiert werden.

[14] realisiert über imgSeek, http://www.imgseek.net/
[15] Nuance – Dragon Naturally Speaking, http://www.nuance.com/dragon/
[16] http://cslr.colorado.edu/beginweb/speech_recognition/sonic_main.html
[17] http://www-i6.informatik.rwth-aachen.de/web/Research/SRSearch_frame.html

Die Annotation des Videomaterials erfolgt also entweder durch Resynchronisation des Präsentationsmaterials mit der Desktopaufzeichnung mittels ICR oder durch Resynchronisation mit dem Audiodatenstrom vermittels ASR. Laut [HLT06] stufen Rezipienten eine Desktopaufzeichnung und die Folien der Präsentation beim Lernen als wichtiger ein als die Aufzeichnung des Dozenten selbst, woraus abzuleiten ist, dass das Anfertigen einer Desktopaufzeichung in Zukunft auch mehr Akzeptanz finden wird.

4 Kollaborative Annotation von Video-Daten

Neben den vielfältigen Möglichkeiten der automatischen Annotation von Multimediadaten, wie sie im vorangegangenen Kapitel besprochen wurden, soll in diesem Kapitel auf eine kollektive Verschlagwortung von Multimediadaten als Ganzem (traditionelles Tagging) bzw. eine synchrone Verschlagwortung von zeitabhängigen Multimediadaten (sequentielles Tagging) näher eingegangen werden.

Unter dem Begriff „Tagging“ wird eine Verschlagwortung verstanden, d. h. die Annotation von Daten (in unserem Falle Multimedia-Daten) mit Begriffen, die den Inhalt oder die Funktion der annotierten Datei markieren [Je95]. Formal ist ein Tag ein Tripel der Form (u, l, r) wobei u für den Benutzer (user), l für das Schlagwort (label) und r für die Ressource stehen. Die Schlagworte können dabei vom Autor der verschlagworteten Ressource selbst, von einem dazu bestimmten Experten, oder aber auch von allen Benutzern (kollaboratives Tagging oder Social Tagging) der Datei gemeinsam vergeben werden.

Aktuelle kollaborative Tagging Systeme wie z. B. delicious[18], bibsonomy[19], My Web 2.0[20] oder das deutschsprachige mister-wong[21] verschlagworten Ressourcen derzeit als Ganzes und sind nicht in der Lage, einzelne Abschnitte dieser Ressource (sofern diese nicht über einen URI identifiziert werden können) gezielt zu annotieren. Man unterscheidet generell zwischen deskriptiven (auch objektiven) Tags, die eine Ressource oder deren Eigenschaften objektiv beschreiben (hierzu zählen inhalts-basierte Tags, kontext-basierte Tags und attributive Tags), und funktionalen Tags, d. h. Tags, deren Bedeutung in der Regel einen ganz bestimmten Zweck anzeigt, der mit der Ressource in Verbindung steht, und der sich meist lediglich dem Tag-Autor allein erschließt und Nutzen bringt (differenziert in subjektive Tags und organisatorische Tags). Siehe [GH06] und [Xu06] für eine detaillierte Übersicht der unterschiedlichen Tag-Kategorien und ihrer Funktion.

Ressourcen jeglicher Art lassen sich vermittels Tags verschlagworten. Diese Schlagworte können dann im Rahmen einer Suche zusätzlich zu den bereits vorhandenen Deskriptoren (Metadaten) genutzt werden. Dabei ist zu beachten, dass kollektives Tagging und die Einbeziehung kollektiv vergebener Tags in die Suche veränderte Rahmenbedingungen für die Suche schaffen, die bereits eingehend untersucht worden sind [Ha06]. Funktionale (subjektiv vergebene) Tags sind in der Regel nur für den Tag-Autor zum Wiederauffinden einer

[18] delicious, http://del.icio.us/
[19] bibsonomy, http://www.bibsonomy.org/
[20] My Web 2.0 http://myweb2.search.yahoo.com/
[21] mister-wong, http://www.mister-wong.de/

verschlagworteten Ressource von Nutzen, während deskriptiv vergebene Tags objektiveren Ansprüchen genügen und auch allgemein für alle in der Suche von Nutzen sind, um neue, bislang unbekannte Ressourcen zu entdecken. Die Verteilung kollektiv vergebener Tags folgt einem Potenzgesetz [GH06], d. h. für eine bestimmte Ressource werden einige wenige Tags sehr oft verwendet, während der Hauptanteil der übrigen Tags für diese Ressource im so genannten „Long Tail“ -Bereich der Tagverteilung liegt, d. h. nur sehr selten vergeben wurde. Diese Eigenschaft kann dazu genutzt werden, zuverlässige Suchergebnisse zu gewinnen bzw. bei Miteinbeziehung der „Long Tail“ -Ergebnisse auf ungeahnte Assoziationen und Querverbindungen zu schließen.

Ein typischer Vertreter einer Suchmaschine mit kollektiv verschlagworteten Multimediadaten ist die bekannte Videosuchmaschine YouTube. Benutzer können dort eigenes Videomaterial einstellen und alle darin vorhandenen Videoclips kollektiv verschlagworten. Kollektive Tags und zusätzlich vom Autor eingegebene Metadaten werden dann bei einer Suche in YouTube in Kombination genutzt. Neben den Suchergebnissen, die durch einen eingegebenen Suchbegriff erzielt wurden, ist YouTube in der Lage, zu einem angezeigten Video anhand der kollektiven Tags weitere ähnliche Videos aus seinem Datenbestand herauszusuchen.

Die kollektive Annotation in der Suchmaschine YouTube oder anderen auf diesem Prinzip basierenden Suchmaschinen (z. B. Google Video oder yahoo! video search[22]) ist stets darauf beschränkt, die vorhandenen Ressourcen als Ganzes zu verschlagworten. Während diese Einschränkung bei zeitunabhängigen Medien nur selten von Nachteil ist – auch wenn ein langes Textdokument als Ergebnis zurückgeliefert wird, kann der Suchbegriff darin leicht mittels einer daran anschließenden Volltext-Suche gefunden werden – kommt dieser Nachteil bei zeitabhängigen Medien voll zum Tragen. Die anschließende Suche innerhalb einer gefundenen Videodatei nach einem bestimmten Suchbegriff gestaltet sich als schwierig. Daher liegt der Schluss nahe, die kollektive Annotation synchron zu einem zeitabhängigen Medium durchzuführen. Zu diesem Zweck wird bei OSOTIS zu jedem vergebenen Tag zusätzlich zum Namen des Nutzers, der das Tag vergeben hat, der Zeitpunkt innerhalb einer Videodatei, zu dem das Tag vergeben wurde, notiert. Diese Art der kollektiven Verschlagwortung bezeichnen wir als synchrones oder sequentielles Tagging. Formal wird das Tripel (u, l, r) also mit einer Funktion $c(r)$ um eine zeitliche Koordinate innerhalb der Ressource erweitert zu $(u, l, c(r))$.

Soll ein Tag nicht nur einen Einzelzeitpunkt sondern ein definiertes Intervall beschreiben, muss jeweils ein Anfangs- und ein Endzeitpunkt zusammen mit dem Tag vermerkt werden. Dieser kann entweder durch den Benutzer selbst oder aber auch durch eine automatische Kontextanalyse bestimmt werden. Die Funktion $c(r)$ kann also auch einen Abschnitt innerhalb einer Ressource beschreiben.

Sequentielles Tagging sowie die automatisierte Resynchronisation des verwendeten Präsentationsmaterials bilden die Basis der Video-Suchmaschine OSOTIS. Die gewonnenen semantischen Annotationen werden als Metadaten parallel zu den Multimediadaten im MPEG-7 Format [CSP01] kodiert. Die Kodierung sequentieller Tags mit Hilfe des MPEG-

[22] yahoo! video search, http://video.search.yahoo.com/

7 Standards wird in [SW06b] näher beschrieben. Aus den MPEG-7 Metadaten wird ein Suchindex aufgebaut, ohne dass ein erneutes Retrieval notwendig ist.

5 OSOTIS – eine kollaborative, inhaltsbasierte Video-Suchmaschine

OSOTIS als Video-Suchmaschine und Web-2.0-Social-Tagging-System hat sich auf die Verwaltung, Annotation und Suche von Lehr- und Lernvideos, und insbesondere von Lehrveranstaltungsaufzeichnungen spezialisiert. Dabei kommen verschiedene Konzepte zum Tragen, um die Recherchierbarkeit der Videodaten mit höherer Feinheit als bisher zu ermöglichen.

OSOTIS verwendet zur Suche sowohl Standard-Suchkriterien, wie z. B. Name des Autors oder andere autorenbezogene Metadaten sowie darüber hinaus eine schlüsselwortbasierte Suche sowohl auf Basis des synchronisierten Präsentationsmaterials als auch mit Hilfe des kollektiven, sequentiellen Taggings. Auf Grund einer Vorabanalyse des textuellen Präsentationsmaterials mit Berücksichtigung von Schriftschnitt und Position in Verbindung mit TF/IDF Metriken[23][PC98] wird die Relevanzgewichtung und damit auch die Qualität der erzielten Suchergebnisse verfeinert. So werden z. B. Videodaten, bei denen das gesuchte Wort in einer Überschrift auftritt, als relevanter eingestuft als Videodaten, bei denen dieses Wort lediglich in einem Nebenkommentar vorkommt. Dies bekräftigt unseren Ansatz, das textuelle Präsentationsmaterial als Grundlage der Schlüsselwörter zu verwenden, da dort der semantische Inhalt des Videos direkt und in kompakter Form niedergeschrieben steht.

OSOTIS präsentiert sich dem Benutzer mit einer einfachen Eingabemaske, in der ein oder mehrere Suchbegriffe eingegeben werden können. Nach inhaltlicher Relevanz wird daraufhin eine Liste mit Suchergebnissen präsentiert und nach Auswahl eines Ergebnisses wird dieses direkt und genau ab der relevanten Stelle wiedergegeben (vgl. Abbildung 2).

Neben der inhaltsbasierten Suche bietet OSOTIS angemeldeten Benutzern die Möglichkeit, das verfügbare Videomaterial mit eigenen sequentiellen (zeitbezogenen) Tags zu annotieren. Auf diese Weise können bestimmte, besonders interessante Abschnitte innerhalb eines Videos besonders hervorgehoben und kategorisiert werden. Eine so genannte „Tag-Cloud“ (siehe Abb. 1, rechts oben) gibt einen Überblick wahlweise über alle aktuell verwendeten Tags und deren Häufigkeit oder gestattet eine nutzer- bzw. mediumbezogene Filterung der angezeigten Tags. Dadurch kann sich der Benutzer auf einen Blick darüber informieren, welche Themen (1) der komplette Videodatenbestand von OSOTIS beinhaltet, (2) ein bestimmtes Video aufweist oder (3) ein bestimmter Nutzer vergeben und annotiert hat. Die in der Tag-Cloud notierten Begriffe selbst können ebenfalls direkt durch einfaches Anklicken zur Suche und Filterung genutzt werden.

Darüber hinaus bietet OSOTIS angemeldeten Benutzern die Möglichkeit, ohne HTML-Kenntnisse eine eigene Webseite zu gestalten, auf der ausgewählte Videos zusammengestellt und präsentiert werden können. So kann der Nutzer z. B. interessante Videos ei-

[23] TF - Term Frequency, IDF - Inverse Document Frequency

Abbildung 2: Suchergebnis für den Begriff „Hieroglyphen“. Es wird dabei angezeigt, an welcher Stelle im Video der Suchbegriff auftritt. Mit einem Klick auf die hervorgehobenen Segmente, wird das Video an dieser Stelle wiedergegeben.

ner Vorlesungsreihe zu eigenen Kollektionen gruppieren. Neben der Vergabe eigener Tags können auch Kommentare und Diskussionen an ausgewählte Video-Positionen „geheftet" werden, in denen mehrere Nutzer den betreffenden Videoausschnitt diskutieren und beurteilen können. Diese Diskussionen erweitern die Annotation und können ebenfalls durchsucht werden.

Das Anmelden von durchsuchbarem Videomaterial bei OSOTIS kann aktuell auf drei unterschiedliche Arten erfolgen: (1) Eigenes Videomaterial kann direkt hochgeladen werden bzw. kann der URL einer oder mehrerer Videodateien direkt angegeben werden. Diese Daten werden nachfolgend direkt durch OSOTIS verwaltet. (2) Videomaterial kann auch über die Angabe der URL einer oder mehrerer Videodateien, die über einem Streaming-Server erreichbar sind, angemeldet werden. OSOTIS lädt diese Daten dann nicht ins eigene System, sondern nutzt lediglich den Link dorthin. Das spart zwar eine redundante Datenhaltung, macht jedoch ein regelmäßiges Überprüfen der betreffenden URLs auf Konsistenz notwendig. (3) Parallel zu den Videodaten kann auch textuelles Präsentationsmaterial[24] hochgeladen werden, das zur automatischen Annotation verwendet wird.

Aktuell (Stand: 05/2007) hält OSOTIS ca. 1700 Videos in englischer und deutscher Sprache vor, von denen ca. 50 % automatisch mit Hilfe des verfügbaren Präsentationsmaterials annotiert worden sind. Der Aufwand der technischen Analyse inklusive der automatischen Annotation benötigt in Abhängigkeit vom vorliegenden Videoformat ca. 3–10 Minuten pro Medienstunde. Das gesamte Videomaterial kann kollaborativ verschlagwortet werden. Aktuell erfolgt dies durch ca. 500 aktive Nutzer. Hierzu ist anzumerken, dass eine aussagekräftige Evaluation der Suchergebnisse von OSOTIS derzeit noch nicht zufriedenstellend durchgeführt werden konnte, da die bislang vorhandene Menge an kollaborativ erstell-

[24] aktuell nur in Form von Adobe PDF- Dokumenten

ten Schlagworten noch zu gering ist. Aktuell werden die an der FSU Jena aufgezeichneten Lehrveranstaltungen wöchentlich in OSOTIS eingestellt und von den Studierenden rege verschlagwortet. Wie für ein Web 2.0 System üblich, wächst der Nutzen des Systems mit der Anzahl der daran aktiv teilnehmenden Benutzer. OSOTIS ist unter dem URL `http://www.osotis.com` frei zugänglich.

6 Zusammenfassung und Ausblick

OSOTIS ermöglicht eine automatische inhaltsbezogene Annotation von Videodaten und dadurch eine zielgenaue Suche auch innerhalb von Videos. Neben objektiv gewonnenen zeitabhängigen Deskriptoren, die über eine automatische Synchronisation von ggf. zusätzlich vorhandenem textuellen Material mit den vorliegenden Videodaten gewonnen werden, können registrierte Nutzer eigene, zeitbezogene Schlagwörter und ganze Kommentare innerhalb eines Videos vergeben, die zur Implementierung einer personalisierten Suche verwendet werden.

Die aktuelle Weiterentwicklung von OSOTIS erstreckt sich neben einer weiteren, qualitativen Verbesserung der damit erzielten Suchergebnisse auf den Bereich des Social Networking und einer Erweiterung des Konzeptes des sequentiellen Taggings. Wie andere Social-Networking-Systeme auch, sollen Benutzer OSOTIS ebenfalls als Kommunikations- und Organisationsplattform nutzen können. So ist z. B. die Bildung von speziellen Lerngruppen angestrebt, die ein gemeinsames Programm an Lehrveranstaltungen absolvieren, diese annotieren, darüber diskutieren und mit Anmerkungen versehen können. Die persönlich vergebenen Tags ermöglichen die Generierung von Nutzerprofilen. Nutzer mit ähnlichen Profilen haben mit hoher Wahrscheinlichkeit ähnliche Interessen oder Expertise. Auf diese Weise lassen sich zuvor ungeahnte Querverbindungen zwischen dem vorhandenen Videomaterial knüpfen und auf Ähnlichkeit basierende Suchfunktionen realisieren. Den Nutzern wird es ermöglicht, eigene Kompetenznetzwerke aufzubauen.

über das zeitbezogene, sequentielle Tagging mit einfachen Schlagwörtern hinaus, werden auch zeitbezogene Annotationen in Form von Diskussionen oder Fragestellung ermöglicht. Dadurch ergeben sich neue Formen der Nutzer-Nutzer-Interaktion, die eine Evaluation der begutachteten Videoinhalte gestatten. Neben der zeitlichen Dimension sollen auch Orts- und Positionsangaben innerhalb eines Videobildes in Form von multidimensionalem Tagging realisiert werden. Auf diese Weise lassen sich spezielle Bildinhalte eines Videos im Rahmen eines bestimmten Beobachtungszeitraumes hervorheben und mit Annotation versehen.

Literaturverzeichnis

[CH03] Y. Chen und W. J. Heng. Automatic Synchronization of Speech Transcript and Slides in Presentation. In Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), Seiten 568–571. Circuits and Systems Society, May 2003.

[CSP01] S. F. Chang, T. Sikora und A. Puri. Overview of the MPEG-7 Standard. IEEE Trans. Circuits and Systems for Video Technology, 11(6):688–695, 2001.

[GH06] S. Golder und B. A. Huberman. The Structure of Collaborative Tagging Systems. Journal of Information Science, 32(2):198–208, April 2006.

[Hu93] X. Huang, F. Alleva, H. W. Hon, M. Y. Hwang und R. Rosenfeld. The SPHINX-II speech recognition system: an overview. Computer Speech and Language, 7(2):137–148, 1993.

[HLT06] C. Hermann, T. Lauer und S. Trahasch. Eine lernerzentrierte Evaluation des Einsatzes von Vorlesungsaufzeichnungen zur Unterstützung der Präsenzlehre. In DeLFI, Seiten 39–50, 2006.

[Ha06] P. Han, Z. Wang, Z. Li, B. Kramer und F. Yang. Substitution or Complement: An Empirical Analysis on the Impact of Collaborative Tagging on Web Search. In Web Intelligence, Seiten 757–760. IEEE Computer Society, 2006.

[ISO05] ISO/IEC 14496-11, Information technology - Coding of audio-visual objects - Part 11 Scene description and application engine, 2005.

[Je95] L. H. Jeng. Organizing Knowledge (2nd ed.), by Jennifer E. Rowley. JASIS, 46(5):394–395, 1995.

[KHE05] S. Kopf, T. Haenselmann und W. Effelsberg. Robust Character Recognition in Low-Resolution Images and Videos. Bericht TR-05-002, Department for Mathematics and Computer Science, University of Mannheim, 04 2005.

[KY96] K. Knill und S. Young. Fast Implementation Methods for Viterbi-based Word-spotting. In Proc. ICASSP '96, Seiten 522–525, Atlanta, GA, 1996.

[NWP03] C. W. Ngo, F. Wang und T. C. Pong. Structuring lecture videos for distance learning applications. In Proceedings of the Fifth International Symposium on Multimedia Software Engineering, Seiten 215–222. IEEE Computer Society, December 2003.

[PC98] J. M. Ponte und W. B. Croft. A Language Modeling Approach to Information Retrieval. In Research and Development in Information Retrieval, Seiten 275–281, 1998.

[Re07] S. Repp, J. Waitelonis, H. Sack und C. Meinel. Segmentation and Annotation of Audiovisual Recordings based on Automated Speech Recognition. In Proc. of 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), Warsaw, Springer, to be published 2007.

[SW06a] H. Sack und J. Waitelonis. Automated Annotations of Synchronized Multimedia Presentations. In Proceedings of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, CEUR Workshop Proceedings, june 2006.

[SW06b] H. Sack und J. Waitelonis. Integrating Social Tagging and Document Annotation for Content-Based Search in Multimedia Data. In Proc. of the 1st Semantic Authoring and Annotation Workshop (SAAW2006), Athens (GA), USA, 2006.

[Sm00] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta und R. Jain. Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell., 22(12):1349–1380, 2000.

[ONH04] J. v. Ossenbruggen, F. Nack und L. Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web, Part 1. IEEE MultiMedia, 11(4):38–48, 2004.

[Xu06] Z. Xu, Y. Fu, J. Mao und D. Su. Towards the semantic web: Collaborative tag suggestions. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May, 2006.

[YOA03] N. Yamamoto, J. Ogata und Y. Ariki. Topic Segmentation and Retrieval System for Lecture Videos Based on Spontaneous Speech Recognition. In Proceedings of the 8th European Conference on Speech Communication and Technology, Seiten 961–964. EUROSPEECH, September 2003.