Hans-Christoph Hobohm (Hrsg.)

Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten

Information Science between Virtual Infrastructure and Material Lifeworlds

Unter Mitarbeit von Judith Pfeffing

Proceedings des 13. Internationalen Symposiums
für Informationswissenschaft (ISI 2013)

Potsdam, 19. bis 22. März 2013

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Ein ganzheitlicher Ansatz zur Digitalisierung und Extraktion von Metadaten in Videoarchiven

Marc Ritter, Robert Herms, Robert Manthey, Maximilian Eibl

Technische Universität Chemnitz, Professur Medieninformatik
D-09107 Chemnitz, Deutschland
{ritm|robeh|mrob|eibl}@informatik.tu-chemnitz.de

Zusammenfassung

Dem ständigen Wachstum von Multimediaarchiven mit Automatisierung zu begegnen erfordert neue Ansätze und Lösungen bei der Extraktion und Verwaltung von Metadaten. Dazu wurden die Frameworks AMOPA und Imtecs entwickelt, um Informationen über den Inhalt der Multimediadaten und die technischen Randbedingungen ihres Einspielvorganges zu extrahieren und den nachfolgenden Stationen des Workflows zur automatischen Steuerung ihrer Arbeitsschritte zur Verfügung zu stellen. Im Archiv angewandt, ermöglichen sie sowohl umfangreichere und komplexere als auch spezifischere Suchen und Anfragen.

Abstract

Multimedia archives have been growing continually throughout the last decades. To meet the need to develop essential new solutions for automated extraction, administration and maintenance of metadata, we propose integrating two frameworks with different tasks to create a unique system with the capability to extract both content-based information and technical boundary conditions during the ingest process. This will make it possible to control and

In: H.-C. Hobohm (Hrsg.). Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten. Tagungsband des 13. Internationalen Symposiums für Informationswissenschaft (ISI 2013), Potsdam, 19.–22. März 2013. Glückstadt: Verlag Werner Hülsbusch, 362–371.

operate the workflows of subsequent tasks. Applied to archives, they will enable more comprehensive, more complex and even more specific search requests to be processed.

1 Herausforderungen

In Archiven sind audiovisuelle Medien zum Teil unerschlossen bzw. liegen in Form von Videobändern vor. Diese durch Beschreibungsdaten (Metadaten) recherchierbar zu machen, entwickelt sich zu einer Herausforderung bezüglich Kapazität und Zeit. Je spezifischer das Metadatenvokabular ist, desto größer sind die Möglichkeiten der Informationsrecherche sowie der Automatisierung von Prozessen (vgl. Hercher et al. 2011). Eine umfassende Erschließung der Inhalte setzt die Sammlung möglichst vieler relevanter Metadaten voraus (Airola et al. 2003).

Diese Herausforderung kann durch einen möglichst frühzeitigen Ansatz, z. B. während des Ingests, bewältigt werden. Ingest bezeichnet das Digitalisieren und Einspielen audiovisueller Inhalte in meist serverbasierten Systemen. Dies bildet die erste Stufe zur Transformation von audiovisuellen Medien in IT-basierte Strukturen. Der automatisierte Ingest stellt einen komplexen Workflow dar, welcher entsprechende Hardware- und Softwarekomponenten voraussetzt (vgl. Borgotallo et al. 2011).

Für den Ingest wurde deshalb ein Framework entwickelt, das eine Kollektion von Komponenten beinhaltet, die in ihrer Gesamtheit eine Middleware zur Automatisierung eines Ingest-Workflows bereitstellt und die Sammlung von Metadaten zu dessen technischen Randbedingungen für die eingespielten Medien erfasst (vgl. Abb. 1, links). Hierzu zählen die eingesetzten Technologien sowie Hard- und Softwarestrukturen. Für den Lebenszyklus audiovisueller Medien bedeutet dies, dass diese technisch umfangreicher dokumentiert sowie für den Endkunden transparenter erscheinen, womit gleichermaßen die Recherchemöglichkeiten gesteigert werden können, da nach Parametern wie z. B. verwendeter Videoschnittstelle oder ursprüngliches Videokassettenformat gesucht werden kann. Zudem ergeben sich für den Workflow zusätzliche Möglichkeiten beispielsweise zur Detektion von Störquellen oder zur automatisierten Fehlerbeseitigung im eingespielten Material, da gegebenenfalls bekannt ist, welcher Kassettentyp oder Videoplayer ein bestimmtes Rauschen verursacht.

Die inhaltliche Beschreibung von audiovisuellen Medien ist eine notwendige Bedingung, um Inhalte in Archiven mit zahlreichen Informationsträgern wiederauffindbar respektive recherchierbar zu gestalten, indem beispielsweise Metadaten in Form von geordneten Suchindizes abgespeichert werden. Im Idealfall reduziert sich die im Allgemeinen aufwendige intellektuelle Annotation somit lediglich auf einen qualitativ orientierten Kontrollschritt nach der automatisierten Inhaltsanalyse, wobei zusätzlich nutzergenerierte Metadaten abgespeichert werden können (vgl. Abb. 1, rechts).

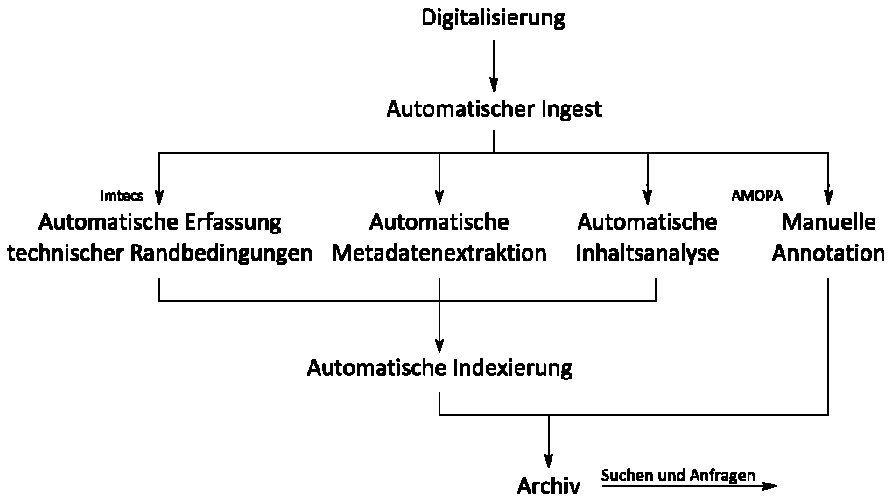

Abb. 1 Das Einspielen analoger Medien beginnt mit der Digitalisierung des Signals und seiner Umwandlung in vorgegebene digitale Datenströme (Ingest), deren Eigenschaften auf ihre spätere Verwendung abgestimmt werden. So benötigen die Automatische Inhaltsanalyse und die Manuelle Annotation meist nur auflösungsreduzierte, einzelbildbasierte Kopien des Videos, die Automatische Metadatenextraktion nur ausgewählte Teile davon und die Automatische Erfassung technischer Randbedingungen nur Informationen über die am Einspielen beteiligten Geräte. Der Datenstrom selbst wird in größtmöglicher Qualität als Videodatei direkt im Archiv abgelegt. Aufbauend auf den automatisch erfassten Metadaten der Technik, des Datenstroms und der Semantik erfolgt die Indexerstellung welche durch Manuelle Annotationen ergänzt wird und somit ein effizientes Suchen und Abfragen des Archivbestandes ermöglicht.

Um diese Bedingungen zu erfüllen und eine ausschlaggebende inhaltliche Beschreibung automatisch zu erhalten, ist eine strukturelle und substanzielle Analyse zur Aufbereitung des Videos notwendig. Zu diesem Zweck erfolgt

u. a. eine akustische Erkennung der Sprecher, des Gesprochenen, sowie weiterer domänenspezifischer multimedialer Entitäten.

Die in Hubbes (2003) vorgestellte Idee einer Middleware für die IT-gestützten Fernsehproduktion wird hierbei aufgegriffen und durch kosteneffiziente, quelloffene und anpassbare Software realisiert. Um einen flexiblen Zugang und einfachen Datenaustausch der Metadaten zu ermöglichen erfolgt deren Verarbeitung, Speicherung und Verteilung im XML- bzw. BMF-Format (Hercher/Mitzscherlich/Sack 2011).

2 Systembeschreibung

Das vorgestellte ganzheitliche System besteht im Wesentlichen aus den in grau unterlegten äußeren Komponenten der mittleren Schicht zur Extraktion und Sammlung unterschiedlicher Informationen in Abbildung 1. Dabei werden einerseits technische Randbedingungen extrahiert, andererseits wahlweise einzeln oder kombiniert über eine automatische Inhaltsanalyse oder intellektuelle Annotation inhaltliche Metadaten hinzugefügt. Der dritte Punkt der automatischen Metadatenextraktion über den verwendeten Codecs, Format, Bildrate, Bitrate, Anzahl mehrerer Audio- und Videospuren etc. liegt nicht im Fokus dieser Arbeit und bleibt vorrangig gängigen Media-Asset-Management-Systemen vorbehalten.

2.1 Ingest-Middleware

Zur automatisierten Extraktion der Metadaten über die technischen Randbedingungen des Ingests, bedarf es einer internen Repräsentation der vorhandenen Geräte und Services in Form von Softwareobjekten. Das Imtecs-Framework (Ingest middleware including extraction of metadata from technical constraints; siehe Abb. 2) stellt hierfür eine Reihe von Softwarekomponenten zur Verfügung. Die Architektur des Frameworks sieht vor, dass Geräte durch ‚Device Objects' und Services durch ‚Service Objects' repräsentiert werden. Ein entsprechender Controller übernimmt im Framework die Steuerung des Ingest-Workflows. Er gibt vor, zu welchem Zeitpunkt welche Geräte bzw. Services zu starten sind, damit ein oder mehrere audiovisuelle Medien erfolgreich eingespielt werden können.

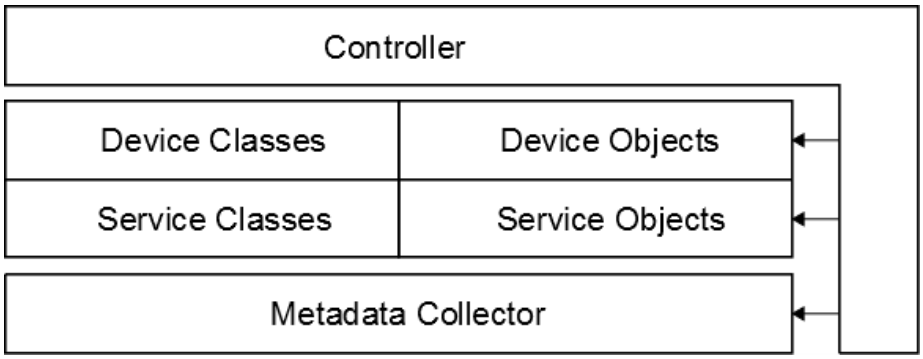

Abb. 2 Architektur des Imtecs-Frameworks

Weiterhin ist im Framework ein Metadata Collector enthalten, der zu einem bestimmten Zeitpunkt, auf Anweisung des Controllers, alle Metadaten der Device Objects und Service Objects sammelt, welche am Ingest-Prozess eines Quellmediums beteiligt sind (siehe Abb. 3).

```
<Ingest_Info version="1.0">
  <Source id="2012-11-17_15-09-01">
    <Medium>Tape</Medium>
    <Format>SVHS</Format>
    <Length unit="minutes">240</Length>
    <Owner>Firma XYZ</Owner>
  </Source>
  <Date>2012-11-17_15-09-01</Date>
  <Duration>04:02:00:00</Duration>
  <Device_List>
    <Player>
      <Name>JVC BR-S522E</Name>
      <Type>SVHS</Type>
      <InterfaceOutput>Component</InterfaceOutput>
    </Player>
    <Converter>
      <Name>Blackmagic Design Mini Converter</Name>
      <Type>Analog to SDI</Type>
      <InterfaceInput>Component</InterfaceInput>
      <InterfaceOutput>SDI</InterfaceOutput>
    </Converter>
    <CaptureDevice>
      <Name>Telestream Pipeline Quad</Name>
      <Type>SD Hardware-Network-Encoder</Type>
      <InterfaceInput>SDI</InterfaceInput>
      <InterfaceOutput>Ethernet</InterfaceOutput>
    </CaptureDevice>
  </Device_List>
</Ingest_Info>
```

Abb. 3 Beispiel für die im Ingest-Workflow extrahierten Metadaten, welche die beim Einspielen einer VHS-Kassette beteiligten Device und Service Objects beschreibt.

2.2 Automatische Inhaltsanalyse

Das Forschungswerkzeug AMOPA (Automated MOving Picture Generator; siehe Abb. 4) ermöglicht die Umsetzung nicht-linearer prozess-gesteuerter Konzepte zur strukturellen und inhaltlichen Analyse audio-visueller Medien (Knauf et al. 2011: 32 f.).

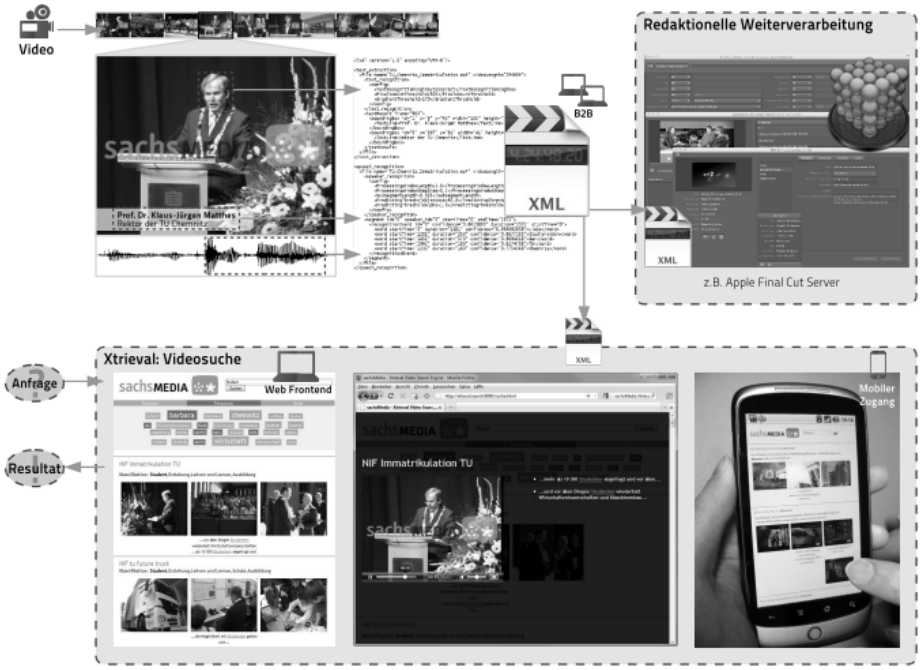

Abb. 4 Workflow zur Nutzung inhaltlicher Metadaten. Links oben: Videodaten werden in Schnitte zerlegt, wobei Gesichter, Textfelder und Sprache extrahiert wird und in XML-Form als Datenbestand (oben Mitte) abgelegt, die im Anschluss einerseits redaktionell weiterverarbeitet werden können (oben rechts). Andererseits kann über ein Suchinterface (unten links) nach extrahierten Metadaten gesucht werden, deren originäre Quellen dann individuell per Browser (unten Mitte) oder über mobile Geräte (unten rechts) abgerufen werden können. (Quelle: sachsmedia [2010])

Die strukturelle Inhaltsanalyse integriert eine Szenenwechselerkennung, welche 98% aller harten und 70% aller weichen Schnitte erkennt (Ritter/Eibl 2009) und mit deren Hilfe Videos über Zeitmarken in verschiedene Segmente unterteilt werden, was eine schnellere Navigation begünstigt, indem nur ein einzelnes repräsentatives Schlüsselbild aus jeder Szene extrahiert und dem Nutzer angezeigt wird.

Die Ermittlung der Videostruktur gestattet gleichermaßen das Auffinden von wiederkehrenden Nachrichtensprechern in Nachrichtensendungen ebenso wie die Identifizierung von Dialogsequenzen bei ständig wechselnden Sprechern. Redundante Sequenzen können somit sogar zusammengefasst und in der Darstellung auf ein Minimum reduziert werden.

Die inhaltliche Analyse komplementiert den vorausgehenden Schritt, indem Objekte in den Schlüsselbildern detektiert und in eine Datenbank abgelegt werden. Beispielsweise selektiert eine Partikelschwarmoptimierung hautfarbenähnliche Flächen im HSV-Farbraum vor (Lang/Ritter 2012), die mit Detektoren, welche auf erweiterten Haar-Merkmalen und dem Verfahren von Viola/Jones (2004) basieren, nach Gesichtsmustern in Echtzeit gescannt werden.

Einen weiteren Baustein bilden die Untersuchung des Audiomaterials sowie die Erkennung von Sprecherwechsel und Sprache. Merkmale werden dazu über die frei verfügbare Software *openSmile* (Eyben et al. 2010) extrahiert, wobei Sprecherwechsel durch die Fenstertechnik mittels Gaußschen Mischverteilungsmodellen identifiziert werden. Zur Erkennung der Sprache wird die Microsoft Speech API verwendet. Dabei wird jedes Wort mit einem Zeitstempel versehen.

Die gewonnenen textuellen Informationen visueller und akustischer Signale bilden eine weit genutzte Basis für nachfolgende Schlagwortrecherchen in Videoarchiven. Obgleich Störquellen oder sogar die inhaltliche Heterogenität audiovisueller Medien zu inkorrekten Beschreibungen führen können, lässt sich aus diesen dennoch ein durchsuchbarer und nutzbarer Index erstellen, indem bei der Suche weitere Nachverarbeitungsfilter und -techniken wie Wörterbücher Anwendung finden.

2.3 Intellektuelle Annotation

Ein zusätzliches Annotationswerkzeug bietet eine grafische Benutzerschnittstelle (siehe Abb. 5), mit deren Hilfe die zuvor extrahierten Schlüsselbilder mit grafischen Elementen und Markierungen versehen und benannt werden können, um so nutzerbasiert Objekte oder Szenen im Videomaterial hervorzuheben, zu redigieren, mit zusätzlichen Informationen zu versehen oder einfach zu beschreiben (Ritter/Eibl 2011).

Abb. 5 Realisiert mittels des Eclipse RCP-Frameworks (McAffer et al. 2005), ermöglicht das Annotationswerkzeug die Nutzung vordefinierter sowie die Erstellung eigener domänenspezifischer Entitäten (bspw. Gesichts- und Körperpartie) wie im linken Arbeitsbereich dargestellt wird. In der Mitte erfolgt die Darstellung der zu annotierenden Szene in Form eines Schlüsselbildes. Der bereits annotierte Kopf wird als blaue Linie hervorgehoben, die in Bearbeitung befindlichen Annotation des Körpers in orange. Im rechten Arbeitsbereich werden zusätzliche Informationen zum Annotationsvorgang angezeigt, worunter auch exakte Koordinaten enthalten sind. Der untere Bereich enthält Informationen zur Gruppierung der zu annotierenden Daten sowie zu deren Bearbeitungsstand.

Das Werkzeug bietet auch die Möglichkeit, die Parameter der beschriebenen Prozessstrukturen aus Abschnitt 2.2 zu verändern und so gegebenenfalls auf unterschiedliche Videomaterialien anzupassen. Gleichzeitig können Zwischenergebnisse sowie Endprodukte der einzelnen Prozesse und Arbeitsschritte visualisiert und interaktiv beeinflusst werden.

3 Zusammenfassung und Ausblick

Die Kombination der beiden Hauptkomponenten Imtecs und AMOPA ermöglicht gleichermaßen die Erfassung technischer Randbedingungen und inhaltlich beschreibender Metadaten von audiovisuellen Medien. Beide können über eine integrierte grafische Benutzerschnittstelle in der Nachverarbeitung modifiziert und durch Export weiteren Programmen verfügbar gemacht werden, um so eine nachfolgende Recherche über Suchanfragen in den erfassten Datenmengen und im Videoarchiv zu erlauben.

Zudem kann sogar der Prozess der intellektuellen Annotation mit dem vorliegenden Annotationswerkzeug von der Verschränkung mit den anderen Schritten profitieren, indem beispielsweise Objekte automatisiert vorselektiert und nur zur Verifikation oder zum Redigieren dem Nutzer vorgestellt werden, womit sich der menschliche Arbeitsaufwand im Vergleich zu einer Brute-Force-Suche innerhalb von Videosequenzen erheblich reduziert.

Darüber hinaus fördert die Kombination dieser Metadatenmengen zugleich umfassendere, korrektere, robustere und somit qualitativ hochwertigere Annotationen, da die Art des Ingest und der weiteren Verarbeitung die Darstellung und Qualität der Medien beeinflussen kann. Diese zusätzlichen Vorkenntnisse erlauben es beispielsweise Darstellungsfehlern, die nur bei bestimmten Quellmaterialien wie VHS auftreten, im Voraus durch entsprechende digitale Aufbereitungsschritte frühzeitig entgegenzuwirken bzw. korrigierend einzugreifen.

Zukünftige Erweiterungen befassen sich einerseits mit der Integration weiterer inhaltlicher Metadaten über die Schlüsselbilder hinaus, beispielsweise durch die Extraktion zusätzlicher Objekte in Form von Gebäuden und Sehenswürdigkeiten. Andererseits sind bei der sprachlichen Analyse erfahrungsgemäß Anpassungen an dialektische Besonderheiten notwendig, um die Erkennungsraten weiter zu verbessern.

Danksagung

Diese Arbeit und das Projekt ValidAX wurden durch das Bundesministerium für Bildung und Forschung (BMBF) gefördert.

Literaturverzeichnis

Airola, D.; Boch, L.; Dimino, G. (2003). Automated Ingestion of Audiovisual Content. RAI Centre for Research and Technical Innovation. http://www.broadcast-papers.com/whitepapers/IBCRAIAutoIngest AVContent.pdf <12.11.2012>.

Borgotallo, R.; Boch, L.; Messina, A. (2011). Automated Industrial Digitization of Betacam tapes – with MXF generation and validation. European Broadcasting Union, Genf. http://tech.ebu.ch/docs/techreview/trev_2011-Q4_betacam-digitization _borgotallo.pdf <12.11.2012>.

Eyben, F.; Wöllmer, M.; Schuller, B. (2010). openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proceedings of ACM Multimedia, ACM MM 2010, 1459–1462.

Hercher, J.; Mitzscherlich, A.; Sack, H. (2011). Bestandsanalyse, Metadaten und Systematisierung. In: Digitalisierungsfibel. Leitfaden für audiovisuelle Archive. Potsdam: transfermedia.

Hubbes, H. (2003). Middleware für den Rundfunk – Was kann sie leisten? In: Institut für Rundfunktechnik, Jahresbericht, 42–47.

McAffer, J.; Lemieux, J.-M. (2005). Eclipse Rich Client Platform: Designing, Coding, and Packaging Java Applications. 2. Aufl., Amsterdam: Addison-Wesley.

Knauf, R.; Kürsten, J.; Kurze, A.; Ritter, M.; Berger, A.; Heinich, S.; Eibl, M. (2011). Produce. Annotate. Archive. repurpose – Accelerating the Composition and Metadata Accumulation of TV Content. In: Evain, J.-P.; Friedland, G.; Sano, M.; Gros, P. (Eds.). Proceedings of the 2011 ACM International Workshop on Automated Media Analysis and Production for Novel TV Services (AIEMPro '11). New York, NY: ACM, 31–36.

Lang, A.; Ritter, M. (2012). Schwarm im Kopf: Gesichtsdetektion mittels Schwarmintelligenz. In: Journal of Junge Wissenschaft 93, 52–59.

Ritter, M.; Eibl, M. (2011). An Extensible Tool for the Annotation of Videos Using Segmentation and Tracking. In: Proceedings of the Design, User Experience, and Usability. Theory, Methods, Tools and Practice. Held as Part of HCI International 2011, Orlando, Florida, July 9–14. Heidelberg: Springer, 295–304.

Ritter, M. (2009). Visualizing steps for shot detection. In: Workshop Information Retrieval 2009 der GI-Fachgruppe Information Retrieval, Held as Part of Lernen Wissen Adaptivität, LWA 2009, Darmstadt, 21.–23. September, 98–100.

sachsmedia (2010). Nutzung automatisch extrahierter Metadaten. http://sachsmedia.tv/files/2010/09/info-sachsMedia-retrieval-framework.pdf <12.11.2012>.

Viola, P. A.; Jones, M. J. (2004). Robust Real-Time Face Detection. In: International Journal of Computer Vision 57 (2), 137–154.