An Ontology-Driven System for the Semantic Analysis of Multimedia Content

Semantic Technologies Conference
June 17, 2009

Walter W. Chang
Senior Computer Scientist

Michael Welch
PhD Candidate, UCLA Computer Science Dept.

Advanced Technology Labs
Adobe Systems, Inc.
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Trends in semantic analysis of multimodal content

- How does one leverage multiple modes of content & metadata?
  - Text documents (formal scripts)
  - Speech-to-Text (STT) transcripts
  - Video Scene Change detectors & Scene Classifiers
  - Face and Object Recognizers, et. al.

Scripts: NL dialog & descriptions
Semantic entities: People, Places, Things
Audio input & STT transcripts
Scene classification
Scene changes & boundaries
Face detection
Using Multimodal Metadata for Video Analysis

- Semantic understanding of video is a difficult problem

- Key approaches:
  - Focus on using reliable lower-level features
  - Develop domain-specific scene content classifiers
  - Leverage ontologies for query

Link to Annotated Bibliography: http://www.cs.ucla.edu/~mjwelch/multimedia/
What types of content semantics can we capture?

- **Video Content**
  - Text Metadata
    - Production Design
      - Set design text
      - Storyboard text
  - Spec & Shooting Scripts
    - Scene List
      - Scene title
      - Scene attributes
      - Scene description
    - Shot List
    - Dialog/Narrative
      - Character dialog
      - Character descriptions
      - Narratives
    - Action Descriptions
  - Audio Metadata
    - Transcripts
    - Automatic Speech-To-Text (STT)
    - Non-dialog features
      - Music
      - Ambient noise
      - Applause
      - Laughter
      - Special EFX
  - Video Metadata
    - Scene Changes
    - Shot Detection
    - Face/Object entities
    - Transitions
    - Generic Sound Library
    - Custom EFX

...and many other content and metadata types
How much multimodal metadata is available?

- Adobe CS can capture a significant quantity of video metadata
- E.g., In “Indiana Jones and The Last Crusade”:

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCENEs</td>
<td>190</td>
</tr>
<tr>
<td>SHOTs</td>
<td>21</td>
</tr>
<tr>
<td>CHARACTERs</td>
<td>42</td>
</tr>
<tr>
<td>DIALOG blocks</td>
<td>861</td>
</tr>
<tr>
<td>DIALOG instr.</td>
<td>145</td>
</tr>
<tr>
<td>TRANSITIONs</td>
<td>22</td>
</tr>
<tr>
<td>ENTITIES</td>
<td>1981</td>
</tr>
<tr>
<td>ENT. REFERENCES</td>
<td>6121</td>
</tr>
<tr>
<td>Temporal Inf.</td>
<td>2933</td>
</tr>
<tr>
<td>STT Timecode Stmts</td>
<td>11098</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------</td>
</tr>
<tr>
<td>CAMERA COORD.</td>
<td>1200</td>
</tr>
<tr>
<td>CAMERA SETTINGS</td>
<td>1200</td>
</tr>
<tr>
<td>LIGHTING DATA</td>
<td>1200</td>
</tr>
<tr>
<td>Other SET metadata</td>
<td>3000</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Triples/TAKE</td>
<td>30K</td>
</tr>
<tr>
<td>Number of TAKEs</td>
<td>10</td>
</tr>
<tr>
<td>Triple Fanout</td>
<td>4</td>
</tr>
<tr>
<td>Running Time</td>
<td>2hr:7min</td>
</tr>
<tr>
<td>Total Triples</td>
<td>1.2 M</td>
</tr>
<tr>
<td>Triples/sec</td>
<td>160</td>
</tr>
</tbody>
</table>
Role of multimodal metadata in multimedia workflows

- What are the key roles for multimodal metadata?
  - Index, search, and organization of content
  - Developing deeper semantic models for video/images:
    - SCENEs and OBJECTs within scene
    - BEHAVIORs and actions of objects within scene
    - Strong entity typing and supertypes / subtypes
  - Computation of “SIMILARITY” measures

- When should multimodal metadata be collected?
  - More opportunities in earlier stage of content production workflow
  - Much more difficult when downstream of metadata creation
    - Video production mixes in sounds EFX, music
    - Original context and workflow state is lost
Role of multimodal metadata in multimedia workflows

- What are the different levels of content search & organization that need to be supported?

  - Search & organize items by:
    - String, synonym & hypernym matching

  - Search & organize by physical concepts:
    - Baseball game scenes
    - Car chase scenes

  - Search & organize by abstract concepts:
    - “Scary” scene
    - “Happy” scenes
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Development of a semantic analysis platform

- What are the key platform elements?
  - Multistage pipeline of text content analysis modules
  - Text feature extractors
  - Text metadata capture and document analysis
  - Metadata & analysis repository
  - Advanced semantic analysis
  - Metadata publishing
    - Search & organization
    - Related content & ad recommendations
  - Domain ontologies
Adobe ATL content intelligence platform (2008)

1. Input media into system
2. Extract, structure, & create text
3. Create semantic metadata & tags
4. Normalize & persist metadata
5. Retrieve, filter, and analyze all metadata
6. (Re-)Use metadata in client applications

Content input
- Upload interface
- Tools & utilities

Text extraction
- Layout extraction
- Page/section segmentation
- Text extraction
- Text glyph filtering
- Stopword filtering
- Term stemming

Metadata generation
- Keyterm entity extractor
- Categorizer & theme analyzer
- Summarizers

Metadata persistence
- XMP metadata services
- Document metadata persistence services

Semantic analysis
- Category & summary filters
- Category taxonomy rule engine
- Adobe keyterm ranker

Essence generation
- Weight categories & themes
- Recommend rule-based categories
- Recommend doc & page Keyterms

Indexes
- XML

Text documents

Indexes
- XML

Tools & utilities

Adobe keyterm ranker

Category & summary filters

Category taxonomy rule engine

Adobe keyterm ranker

Weight categories & themes

Recommend rule-based categories

Recommend doc & page Keyterms

Essence generation

Metadata Repository

Summarizers

Metadata persistence services

Document metadata persistence services

XMP metadata services

Keyterm entity extractor

Categorizer & theme analyzer

Summarizers

Metadata generation

Metadata persistence

Retrieve, filter, and analyze all metadata

(Re-)Use metadata in client applications

Content input

Text extraction

Metadata generation

Metadata persistence

Semantic analysis

Essence generation

1. Input media into system
2. Extract, structure, & create text
3. Create semantic metadata & tags
4. Normalize & persist metadata
5. Retrieve, filter, and analyze all metadata
6. (Re-)Use metadata in client applications
Extending the content intelligence platform for multimedia

- What are the key platform elements?
  - Multistage pipeline of **text, audio, & video** content analysis modules
  - **Text, audio, and video** feature extractors
  - Multimedia metadata capture and content analysis
  - Multimedia metadata & analysis repository
  - Advanced **multimodal** semantic analysis
  - Metadata publishing
    - Search & organization
    - Related content & ad recommendations
  - More ontologies
Adobe ATL multimedia content intelligence platform

1. Input media into system
   - Upload interface
   - Tools & utilities
   - Video, audio, frame, & cc capture
   - Audio text transcription generator
   - Upload interface

2. Extract, structure, & create text
   - Layout extraction
   - Page/section segmentation
   - Text extraction
   - Text glyph filtering
   - Stopword filtering
   - Term stemming
   - Audio text transcription XML parser
   - Adobe movie script parser

3. Create semantic metadata & tags
   - Keyterm entity extractor
   - Categorizer & theme analyzer
   - Summarizers
   - Natural Language processor
   - Named entity recognizer/extractor
   - Script entity recognizer/extractor

4. Normalize & persist metadata
   - XMP metadata services
   - Document metadata persistence services
   - Metadata Repository RDF-Triple store
   - Metadata persistence
   - AV metadata persistence services
   - XMP Metadata services
   - Script entity filters & analyzer
   - Video index & report generation

5. Retrieve, filter, and analyze all metadata
   - XMP metadata services
   - Metadata Repository RDF-Triple store
   - Script scene filters & analyzer
   - Video entity indexing

6. (Re-)Use metadata in client applications
   - Script entity filters & analyzer
   - Video entity report generators
   - Recommend doc & page Keyterms
   - Recommend rule-based categories
   - Adobe keyterm ranker
   - Weight categories & themes

Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Capturing and representing content semantics

- Movie scripts are a rich source of metadata
  - SCENEs provide semantic structure and location information
  - SHOTs provide units of context and action
  - ACTIONs describe semantic progression of events
  - TRANSITIONS describe camera actions (e.g., FADE IN)
  - CHARACTERs indicate source of activities
  - DIALOG elements provide character communication

- Challenges:
  - Spec script and shooting script formats vary
  - Scripts contain semi-structured and unstructured data

- Approach: develop platform to import & capture scripts
Capture rich semantic metadata from movie scripts

Movie Spec.
Script

Adobe Story - Movie Script Capture and Editing Tool
Analyze semantic content for fine-grain entities & actions

- Scripts contain semi-structured content:
  - SCENE headings
    - INTerior, EXTerior,
    - time, place, other attributes
  - SHOT descriptions
    - Shot type, actors or scenes

- Scripts contain unstructured content:
  - ACTION description elements
  - DIALOG elements

- Goal: extract additional fine-grain semantics
Capture fine-grain metadata from movie scripts

Adobe Story
Client Script Application

Script Converters
- Layout analyzer
- Raw script parser
- XML output generator

Script Parsers
- Extractor1
- Extractor2
- Extractor3
- Extractor4
- Extractor5

XML

Movie Script

Metadata Repository

Adobe Script Services

Speech-to-Text Services
- STT LM Creation
- STT Parallel Transcription

Metadata Time Sync. Services
- Script / Video Alignment
- Video Indexing

Natural Language Engines
- Text stream filter
- Speech Taggers
- Entity Extractors
- Lexicon Rules
- Ontology Mgr

Fact Table (Triple Store)

Other non-text Metadata

Image Metadata

Time Sync. Services

Time
- Characters
- Action
- Dialogue
- Objects

Video Indexing

STT Parallel Transcription

Capture rich semantic metadata from STT dialog

- Spec scripts have limited video temporal data
- Audio is another channel of content and metadata
- Speech-to-text is semantically rich, but “noisy”
Large bias problem with default STT language model (LM)

Word feature Prior Prob.
Iran 1.39e-04
Clinton 1.90e-04
baseball 6.55e-05
beach 2.09e-05
gun 6.701497e-05
Solution: build ontology of custom STT lexicons and LMs

News corpus

AdobeFilms script genre taxonomy
Speech-to-text language model
Version 1.0 - June 28, 2008

Genre schema

Film genre taxonomy

Default Language Model M0

Custom Language Model M1

Custom Language Model M2

Custom Language Model M3
Preprocess audio signal via dialog classification

Video
Audio content
Audio signal ontology

Acoustic model / sound classifier

Classifier Segmenter/STT Engine

Machine learning

dialog, music, dialog+music, silence

Content Segment List

<table>
<thead>
<tr>
<th>Time Segment</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01 – T03</td>
<td>music</td>
</tr>
<tr>
<td>T03 – T04</td>
<td>silence</td>
</tr>
<tr>
<td>T05 – T06</td>
<td>dialog</td>
</tr>
<tr>
<td>T06 – T07</td>
<td>silence</td>
</tr>
<tr>
<td>T07 – T08</td>
<td>dialog</td>
</tr>
<tr>
<td>T08 – T09</td>
<td>music</td>
</tr>
<tr>
<td>T10 – T11</td>
<td>silence</td>
</tr>
<tr>
<td>T11 – T12</td>
<td>dialog</td>
</tr>
<tr>
<td>T12 – T13</td>
<td>music</td>
</tr>
<tr>
<td>T13 – T14</td>
<td>silence</td>
</tr>
<tr>
<td>T14 – T15</td>
<td>music</td>
</tr>
<tr>
<td>T15 – T16</td>
<td>dialog</td>
</tr>
<tr>
<td>T16 – T17</td>
<td>silence</td>
</tr>
<tr>
<td>T17 – T18</td>
<td>dialog</td>
</tr>
<tr>
<td>T18 – T19</td>
<td>dialog</td>
</tr>
<tr>
<td>T19 – T20</td>
<td>silence</td>
</tr>
<tr>
<td>T20 – T21</td>
<td>dialog</td>
</tr>
<tr>
<td>T21 – T22</td>
<td>dialog</td>
</tr>
<tr>
<td>T22 – T23</td>
<td>silence</td>
</tr>
<tr>
<td>T23 – T24</td>
<td>silence</td>
</tr>
<tr>
<td>T24 – T25</td>
<td>dialog</td>
</tr>
<tr>
<td>T25 – T26</td>
<td>music</td>
</tr>
</tbody>
</table>
Generate STT transcription using multicore system

Dialog Segment Descriptor Job Queue

Top of Queue
- DSD T17-T18
- DSD T18-T19
- DSD T20-T21
- DSD T21-T22
- DSD T24-T25

Audio File

CPU 01
- DSD T11-T12

CPU 02
- DSD T05-T06

CPU 03
- DSD T15-T16

CPU 04
- DSD T07-T08

Partial Trans. File 1

PTF Merge()

Full STT Transcription File
Generalize capture and representation of content features

- Media content analyzers collect feature metadata
- Feature metadata is probabilistic
- Introduces multimodal feature ambiguity
Generalize capture & representation of content features

- Media content analyzers collect feature metadata
- Feature metadata is probabilistic
- Introduces multimodal feature ambiguity, now what?

Multimodal Metadata Feature Repository

- (SCENE1, stadium, 1.0)
- (bat, mammal, 0.30)
- (bat, baseball, 0.13)
- (diamond, gem, 0.45)
- (diamond, baseball, 0.02)
- (bases, 0.55)
- (basis, 0.27)
- (stick, 0.06)
- (baton, 0.03)
- (baseball bat, 0.01)
- (golfcourse, 0.15)
- (baseball diamond, 0.08)
- (Terence, 0.19)
- (Ray, 0.07)
- (Amy, 0.03)
Develop a multimodal inferencing model

- Automatically find best hypothesis about a scene
  - Movie SCENE is about baseball?
  - Object in SCENE is a baseball diamond?

- Initial approach: Use Bayesian inferencing:

\[
P(H|E) = \frac{P(E|H)P(H)}{P(E)}
\]

where:
- \( H \) = hypothesis,
- \( P(H) \) = prior probability of \( H \) before evidence \( E \)
- \( P(E|H) \) = conditional probability of seeing \( E \) if \( H \) is true
- \( P(E) \) = marginal probability of \( E \) (seeing \( E \) under all \( H[i] \))

\[
P(H|E_1 \land E_2 \land E_i) = \frac{P(E_1|H)P(E_2|H)P(H)}{P(E_1|H)P(E_2|H)\ldots P(E_i|H)P(H) + P(E_1|\sim H)P(E_2|\sim H)\ldots P(E_i|\sim H)P(\sim H)}
\]
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Developing and leveraging heterogeneous ontologies

- Multiple knowledge structures needed for each metadata analyze/extractor for each mode
  - Domain attributes: name, context, description
  - Training corpus for ML
  - Feature schema
  - Feature “vocabulary”

- Build canonical classifier ontology
Adobe ATL Heterogeneous Ontology Framework

Audio content
- Linguistic analyzers
- Entity extractors
- Audio classifiers
- STT extractors
- Face detectors
- Scene detectors
- Video edit toolset

Video content
- STT extractors
- Multimodal metadata interchange & publishing system
- Multimodal query engine
- Semantic metadata storage manager
- RDF Triplestore
- Relational Data store

Ontologies:
- Document ontology
  - Domain schema
  - Scene/Obj schema
  - Ref training corpus
- Audio ontology
  - Genre schema
  - Language model
  - Ref training corpus
- Face detector ontology
  - Facial schema
  - Facial attr rules
  - Ref training corpus
- Scene detector ontology
  - Scene schema
  - Scene attr rules
  - Ref training corpus
- Video ontology
  - Clip schema
  - Visual property model
  - Ref training corpus

Inferencing rules
- Ref training corpus
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Objective: Select suitable entity ID strategy

- **What are good entity identifiers?**
  - “GUIDs” today are overloaded, not consistent: unify in repository
  - Recommendation: use “urn:uuid:” and in particular, RFC 4122 Version 4 to remove IP and timestamp semantics
  - Don’t try to conserve entity IDs
  - Defer entity ID resolution
  - Assign new IDs to entities in statements as encountered e.g., (P,S,O) triples
  - Reconcile entities by inferences when needed (late entity ID binding)

```
ID (Identifier)
  |-- URI (Universal Resource Identifier, RFC 3986)
  |    |-- UUID (Universally Unique Identifier, RFC 4122)
  |         |-- Version 1 (MAC address + Timestamp)
  |         |-- Version 2 (DCE Security, POSIX local domain + user)
  |         |-- Version 3 (MD5 hash of URL, fully qual. domain name, OID)
  |         |-- **Version 4** (Random number)
  |         |    |-- GUID (Globally Unique Identifier)
  |         |    |-- Version 5 (SHA-1 hash)
```
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Semantic representation of movie scripts

- Movie is described by script

- Scripts contain:
  - Semantic Entities
  - Semantic Relationships between entities
  - Structured sequence of semantic actions
  - Rules & constraints for how script can be structured

- Semantic structure of script can be described by:
  - Object model
  - Entity-Relationship model
  - UML Class and Process Model
  - Ontology
Simplified E-R & Relational Model of Movie Script

- **Scripts**: Doc_id, Script_name
- **Locations**
  - Doc_id
  - Loc_id
  - Loc_Type
  - Loc_Title
- **Scenes**
  - Doc_id
  - Scene_id
  - Scn_Type
  - Scn_Name
  - Scn_TOD
- **Objects**
  - Doc_id
  - Obj_id
  - Obj_Name
  - Obj_type
- **Descriptions**
  - Doc_id
  - Action_id
  - Action_Type
  - Action_Desc
- **Dialog**
  - Doc_id
  - Diag_id
  - Dialog_Type
  - Dialog_Text
- **Characters**
  - Doc_id
  - Char_id
  - Char_Name
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Why integrate temporal and semantic metadata?

- Accurate dialog and keyword search of video content
- Automatic scene time-indexing using script semantics
- Fine-grain scene entity classification and time-indexing
- Automated inferencing about video semantics
  - Search by scene entity supertypes
  - Search by scene entity part-hierarchies

- Semantic analysis and time alignment metadata stored into:
  - XMP (see XMP temporal metadata spec.)
  - Adobe metadata repository
Results of STT & SCRIPT time-alignment
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Enabling new multimodal applications

- Time alignment of entities and events within content
- Content search and indexing
- Content (re-)organization
- SCENE & SHOT understanding for planning, editing
- Intelligent recommendation systems for:
  - related video content
  - relevant ads
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications

- Lessons learned
- Next Steps
Lessons Learned

- Capture & represent content semantics into E-R model
  - Analyze & trade-off tension between relational and RDF-Triples
  - Templatize mapping of relational data stores to RDF-Triplestores
  - Identify inference use cases for graph models

- Develop and leverage heterogeneous ontologies
  - Extend ontologies for tracking machine learning corpora
  - Adopt multifaceted ontology approach

- Defer binding of entity & metadata identification

- Incorporate logic-based and statistical inferencing methods for multimodal metadata
Overview

- Trends in semantic analysis of multimodal content
- Role of multimodal metadata in multimedia workflows
- Development of a semantic analysis platform for
  - Capturing and representing content semantics
  - Developing and leveraging heterogeneous ontologies
  - Uniform semantic metadata identification
  - Leveraging relational data stores and RDF-Triplestores
  - Integrating temporal and semantic metadata
  - Enabling new multimodal applications
- Lessons learned
- Next Steps
Next Steps

- Deploy semantic analysis functions into Adobe CS5
- Develop & prototype multimodal query model
- Extend base ontologies
- Questions & comments welcome!

wachang@adobe.com
mwelch@adobe.com