An Ontological Approach to Semantic Video Analysis for Violence Identification

T. Perperis\(^1\) S. Tsekeridou\(^2\)

\(^1\)University of Athens, Department of Informatics and Telecommunications, Greece, (http://www.di.uoa.gr)

\(^2\)Athens Information Technology, Greece, (AIT, http://www.ait.edu.gr)

AIT Research Seminar 2011
AIT, Athens, Greece, 15 February, 2011
Outline

1. Introduction
2. Existing Approaches
3. Ontological Methodology for Violence Identification
4. Conclusions
Outline

1. Introduction
2. Existing Approaches
3. Ontological Methodology for Violence Identification
4. Conclusions
Motivation

<table>
<thead>
<tr>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential growth of Multimedia Content</td>
</tr>
<tr>
<td>Uncontrollable dissemination of Objectionable Content</td>
</tr>
</tbody>
</table>

Common users and Industry:

<table>
<thead>
<tr>
<th>Demand for Intelligent, Human Like methods to:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Automatically Search and Classify video data</td>
</tr>
<tr>
<td>2. Automatically Detect and Annotate dangerous content</td>
</tr>
<tr>
<td>3. Filter out this content, thus enabling high level parental control</td>
</tr>
</tbody>
</table>
**Semantic Gap**

*Main Obstacle*

*Semantic Gap*: Inability of machines to automatically grasp high level semantic concepts from multimodal data.
Bridging The Semantic Gap - Technical Difficulties

Technical Difficulties

1. Audio Analysis
   - Development of efficient audio events detectors

2. Visual analysis
   - Development of efficient algorithms for Shot boundary, Event, Object, editorial effects detection

3. Content Annotation
   - Creation of Multimedia and Semantics metadata descriptions (MPEG-7, ontologies)
   - Interoperability of Multimedia with Domain Ontologies and Low level algorithms

4. Content and Metadata delivery
   - MPEG-21, MPEG-4 etc.

5. Content filtering (user side)
   - MPEG-7, OWL, Ontology Quering (SPARQL, SQWRL)
## Technical Difficulties

1. Audio Analysis
   - Development of efficient audio events detectors

2. Visual analysis
   - Development of efficient algorithms for Shot boundary, Event, Object, editorial effects detection

3. Content Annotation
   - Creation of Multimedia and Semantics metadata descriptions (MPEG-7, ontologies)
   - Interoperability of Multimedia with Domain Ontologies and Low level algorithms

4. Content and Metadata delivery
   - MPEG-21, MPEG-4 etc.

5. Content filtering (user side)
   - MPEG-7, OWL, Ontology Quering (SPARQL, SQWRL)
### Bridging The Semantic Gap - Conceptual Difficulties

#### Elementary Question

*What kind of content is considered harmful?*

#### Answer

*Violence and Pornography*

#### Definition

- **Violence**: Any action or situation, that may cause pain, suffering, physical or mental harm to one or more persons, injury to animals, fire, explosion or destruction of non-living objects.

- **Pornography**: Any situation containing molestation, nudity or sexual actions.
Bridging The Semantic Gap - Conceptual Difficulties

Elementary Question
What kind of content is considered harmful?

Answer
Violence and Pornography

Definition
Violence  Any action or situation, that may cause pain, suffering, physical or mental harm to one or more persons, injury to animals, fire, explosion or destruction of non-living objects

Pornography  Any situation containing molestation, nudity or sexual actions
Outline

1. Introduction
2. Existing Approaches
3. Ontological Methodology for Violence Identification
4. Conclusions
Audio Visual Analysis

Most Approaches Follow the Pattern Recognition Paradigm

1. What to classify?
   - Objects (i.e. Shot, Frame, Bounding Boxes)
   - Features (i.e. Statistics of signal’s properties)

2. Which categories to classify to?
   - Binary (i.e. violence vs non violence)
   - Multiclass (i.e. fights, gunshots, screams)

3. How to classify?
   - Single-Modal vs Multi-Modal
   - Early fusion vs Late fusion
   - Supervised vs Unsupervised
   - Specific Algorithm
Audio Visual Analysis

Most Approaches Follow the Pattern Recognition Paradigm

1. What to classify?
   - Objects (i.e. Shot, Frame, Bounding Boxes)
   - Features (i.e. Statistics of signal’s properties)

2. Which categories to classify to?
   - Binary (i.e. violence vs non violence)
   - Multiclass (i.e. fights, gunshots, screams)

3. How to classify?
   - Single-Modal vs Multi-Modal
   - Early fusion vs Late fusion
   - Supervised vs Unsupervised
   - Specific Algorithm
Audio Analysis

- Significantly easier than visual analysis
- Valuable hints on the identification of gunshots, explosions, screams, music, speech
- Fairly limited approaches for violence detection
  - MoCA: [MoCA, 96]
  - Giannakopoulos et al.: [Giannakopoulos 06,07,08]
- Missing approaches for pornography detection
Visual Analysis

- Richer set of semantics
- Much more challenging task
  - Shot Boundary Detection
  - Object recognition, localisation, tracking
  - Camera movements - Editorial effects
  - Compression standard used
- High Activity in short shots: [Vasconelos, 97], [Lehane, 04]
- Limps Orientation: [Datta, 02]
- Gunshot detection through luminance differences: [Makris, 10]
- Skin detection and motion patterns for pornography: [Jansohn, 09], [Kim, 08]
Multi-Modal Fusion

Violent or pornographic actions rarely appear in isolation (i.e. only in the one of the modalities), the potential of multimodal approaches towards increasing the detection accuracy or the level of extracted semantics has to be explored.

**Early Fusion**

Single modality features are combined in a unique multimodal representation, feeding machine learning algorithms to extract combined semantics ([Jian, 09], [Gong, 08])

**Late Fusion**

Extracted audio and visual mid-level semantics are coupled to achieve higher level of abstraction and improve semantics extraction accuracy ([Nam, 98])
Ontological Approaches?
Ontology Definition

An explicit specification of a conceptualisation [Gruber, 93]

Explicit Specification: An **Ontology** is written using logic or other formal language (Description Logic)

Conceptualization: An **Ontology** is a semantic structure (i.e. tree, graph, hierarchy) which encodes the rules constraining the structure of the domain
Ontology Development II

An ontology is an engineering artifact:
- Constituted by a specific vocabulary describing a certain reality, plus
- A set of explicit assumptions regarding the intended meaning of the vocabulary

An ontology describes a specification of:
- A shared understanding of the domain of interest
- A formal and machine manipulable model of the domain of interest
Ontology Components

Term Vocabulary specification by means of:
- Objects, Concepts, Individuals, Properties
- i.e. Movie, Shot, Frame, hasFrames, hasMovieName

Term meaning specification through:
- Object relationships (i.e. shot has frames, shot is a temporal segment)
- Attributes and properties (i.e. movie has name, frame has number)
- Constraints (i.e. each temporal segment has only one starting time and only one ending time)
Ontology Languages

- **OWL-Lite**: Classification hierarchy and simple constraints
- **OWL-DL**: Maximum expressiveness and practical reasoning algorithms
- **OWL-Full**: Does not enforce a strict separation of classes, properties, individuals and data values
Ontology Formalism - Description Logics

What Are Description Logics? ([Baader, 03])

- A family of logic based Knowledge Representation formalisms
- Decidable fragments of First Order Logic
- Describe domain in terms of classes, relationships and individuals
- Provide inference services

DL Knowledge Base

- ABOX (Assertional Box) → Schema
- TBOX (Terminological Box) → Data
### Ontology Extensions I

#### Semantic Web Rule Language (SWRL)

- A proposal to combine ontologies (OWL-DL) and rules (Rule-ML)
- The standard Horn like rule language of the Semantic Web
- Increase OWL Expressivity
  - Can reason about OWL instances (individuals) in terms of OWL classes and properties
  - In OWL it is not possible to establish that a shot is part a movie, only that the shot is a temporal segment
- Rules are of the form of an implication between an antecedent (body) and consequent (head) - *If body rules hold head rules also hold*
- SWRL reasoning occurs on Instance level
Ontology Extensions I

Semantic Web Rule Language (SWRL)

- A proposal to combine ontologies (OWL-DL) and rules (Rule-ML)
- The standard Horn like rule language of the Semantic Web
- Increase OWL Expressivity
  - Can reason about OWL instances (individuals) in terms of OWL classes and properties
  - In OWL it is not possible to establish that a shot is part a movie, only that the shot is a temporal segment
- Rules are of the form of an implication between an antecedent (body) and consequent (head) - *If body rules hold head rules also hold*
- SWRL reasoning occurs on Instance level
Ontology Extensions II

Semantic Query-Enhanced Web Rule Language (SQWRL)

- SWRL-based language for querying OWL ontologies
- SQL-like operations to retrieve knowledge from OWL
Ontological Artifacts on Semantic Video Analysis

Constraint Domains
- News, Sports, Medical, Surveillance

Application On Violence Detection?
- Video Event Representation Language for Surveillance applications ([VERL, 05])
- Action representation with Context Free Grammars ([Ryoo, 06])
Ontological Methodology

**Goals**

1. Automatically detect any violence hidden in video data
2. Automatically annotate them accordingly
3. Enable content filtering for parental control
4. Not to devise high quality low level analysis processes
5. Combine existing single modality low to mid-level semantics detectors with ontologies and reasoning
6. Open/Extendable framework
Ontological Methodology

Goals

1. Automatically detect any violence hidden in video data
2. Automatically annotate them accordingly
3. Enable content filtering for parental control
4. Not to devise high quality low level analysis processes
5. Combine existing single modality low to mid-level semantics detectors with ontologies and reasoning
6. Open/Extendable framework
Ontological Framework

- **Introduction**
- **Existing Approaches**
- **Ontological Methodology For Violence Identification**
- **Conclusions**

**Ontological Framework**

- **Audio Analysis**
  - Gunshots / Fights / Screams Music / Speech / Others Detection
- **Visual Analysis**
  - Activity / Gunshot Detection
- **Preprocessing**
  - Audio Segmentation
  - Shot Boundary Detection
- **Clustering**
  - Scene Detection
- **Instantiated Ontologies (ABOX)**
- **Ontology Classification** (Pellet)
  - SWRL - 1st Set (Jess)
  - SWRL - 2nd Set (Jess)
- **Harmful Content Ontology**
- **Ontology Classification** (Pellet)
- **Video Structure / Visual / Audio Ontologies**
- **5-Step Knowledge Generation / Inferencing Process**
Major processes of the system

- Preprocessing
- Visual analysis
- Audio analysis
- Scene Detection

- Video Structure Ontology
- Visual Ontology
- Audio Ontology
- Video Structure Ontology

- Inferencing Procedure
- Harmful Content Ontology
Preprocessing
### Preprocessing

#### Segmentation Role
- Define the temporal annotation units
- Feed low level analysis with temporal segments of predefined duration
- Preserve a common time reference and extract sequence and overlapping relations

#### Segmentation
- Initial stream segmentation into fixed duration segments (1-sec as the minimum event duration)
- Shot boundary detection through *Local Content Adaptive Thresholding*
Shot boundary detection

Local Content Adaptive Thresholding

Compute on a 5-D temporal window:

- Local color histogram based interframe differences
- Their mean values
- If the current difference is maximum in the window and is greater than twice the mean window value, a shot cut is detected
Segmentation Semantics - Video Structure Ontology

Video Structure Ontology Captures:

- Authoring Information
- Segmentation Semantics

Video Structure Ontology Comprise:

- Main interconnection point with other ontologies
- The first to be instantiated
Video Structure Ontology

- VSO:MultimediaDocument
- VSO:Movie
- VSO:Scene
- VSO:TemporalSegment
- VSO:LogicalSegment
- VSO:AudioVisualSegment
- VSO:Shot
- VSO:Frame
- VSO:StructuralSegment

Isa relationships:
- VSO:MultimediaDocument is a VSO:Movie
- VSO:Movie hasNextScene, hasPreviousScene
- VSO:Movie hasScenes
- VSO:Scene hasCorrespondingAudioVisualSegments
- VSO:Scene hasNextScene, hasPreviousScene
- VSO:Scene hasCorrespondingTemporalSegments
- VSO:Scene hasCorrespondingStructuralSegments
- VSO:TemporalSegment hasNextSegment, hasPreviousSegment
- VSO:TemporalSegment hasCorrespondingAudioVisualSegments
- VSO:TemporalSegment hasCorrespondingFrames
- VSO:AudioVisualSegment hasCorrespondingAudioVisualSegments
- VSO:Shot hasCorrespondingFrames
- VSO:Frame hasNextSegment, hasPreviousSegment
Visual Analysis / Semantics
# Visual Analysis I

## Activity Detection
- No Activity
- Normal Activity
- High Activity

## Gunshot Detection
- No Gunshot
- Gunshot

## Person Detection
- Haar Based Face Detection
- RGB Skin Detection

---

1[Makris, 10]
Visual Analysis II

Correlation between Violence and Activity classes.
Visual Analysis III

### Visual Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>AM</em></td>
<td>Average overall motion calculated using motion vectors.</td>
</tr>
<tr>
<td><em>MOV</em></td>
<td>Variance of the motion vectors orientations.</td>
</tr>
<tr>
<td><em>OTD</em></td>
<td>Average degree of overlap of the detected people.</td>
</tr>
<tr>
<td><em>MLD</em></td>
<td>Maximum Luminance Difference.</td>
</tr>
<tr>
<td><em>MLI</em></td>
<td>Maximum Luminance Interval.</td>
</tr>
</tbody>
</table>

### Classification Approach

- Weighted kNN classifier for Activity Detection (*AM*, *MOV*, *OTD*)
- Weighted kNN classifier for Gunshot detection (*MLD*, *MLI*)
Visual Ontology I

Include the hierarchical definition of:

- Detected visual objects possibly related with violence (or not)
- Detected visual events possibly related with violence (or not)
- An extensive range of various visual events and objects to provide attachment points for other concept detectors
Visual Ontology II - Objects
Visual Ontology III - Events

- MVisO: VisualEvent
  - isa
  - MVisO: Explosion
  - isa
  - MVisO: Activity
    - isa
    - MVisO: HighActivity
    - isa
    - MVisO: NormalActivity
  - isa
  - MVisO: Gunshot
    - isa
    - MVisO: NoActivity
Audio Analysis / Semantics
# Audio Analysis I²

## Detected Non Violent Classes
1. Music
2. Speech
3. Others1 (Sounds of low energy - Smooth enviromental sounds like backround noise, rain, wind)
4. Others2 (Sounds of abrupt changes in signal energy - Sharp enviromental sounds like thunder, door closing etc.)

## Detected Violent Classes
1. Gunshots
2. Screams
3. Fights

²[Giannakopoulos, 07]
## Audio Analysis II

### Employed Features

<table>
<thead>
<tr>
<th></th>
<th>Frame Feature</th>
<th>Sequence Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Spectrogram</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>2</td>
<td>Chroma 1</td>
<td>$\mu$</td>
</tr>
<tr>
<td>3</td>
<td>Chroma 2</td>
<td>median</td>
</tr>
<tr>
<td>4</td>
<td>Energy Entropy</td>
<td>$\max$</td>
</tr>
<tr>
<td>5</td>
<td>MFCC 2</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>6</td>
<td>MFCC 1</td>
<td>$\max$</td>
</tr>
<tr>
<td>7</td>
<td>ZCR</td>
<td>$\mu$</td>
</tr>
<tr>
<td>8</td>
<td>Sp. RollOff</td>
<td>median</td>
</tr>
<tr>
<td>9</td>
<td>Zero Pitch Ratio</td>
<td>$-$</td>
</tr>
<tr>
<td>10</td>
<td>MFCC 1</td>
<td>$\max/\mu$</td>
</tr>
<tr>
<td>11</td>
<td>Spectrogram</td>
<td>$\max$</td>
</tr>
<tr>
<td>12</td>
<td>MFCC 3</td>
<td>median</td>
</tr>
</tbody>
</table>
Audio Analysis III

One Vs All Strategy

- Split the main problem into K (K = 6) binary sub-problems (e.g. gunshot vs. no gunshot)
- Randomly split the 12-D feature vector into 3, 4-D feature sub vectors
- Train 3 kNN binary classifiers for each subproblem
- Feed the results of the 3 kNN classifiers into a Bayesian Network and produce the probability estimation for the subproblem
- 7 probability estimations are produced
- The winner is the class with the highest probability
## Audio Ontology

**Defines Taxonomy of:**

1. Audio Events (e.g. MSO:Gunshot, MSO:Screams, MSO:explosions, MSO:Speech, MSO:Music)
2. Type of background music (MSO:Pop, MSO:Rock, MSO:Classic etc)
3. Emotional Speech (e.g. MSO:Anger, MSO:Fear, MSO:Cry)

**Contains:**

- A much broader set of audio semantics for future extension
Audio Ontology

MSO:WeaponRelatedSound

MSO:FightRelatedSound

MSO:EnvironmentalSound

MSO:SharpEnvironmentalSound

MSO:SmoothEnvironmentalSound

MSO:PersonRelatedSound

MSO:Scream

MSO:Speech

MSO:Music

MSO:MusicGenre

MSO:hasGenre

MSO:hasEmotion
Scene Detection
Scene Detection

**Shot Clustering**

- Create a 10-D vector composed of Audio And visual classes probability estimations
- Feed a MCL algorithm that uses Random Walks to create clusters
  - Instantiate VSO:Cluster class in the VS Ontology
  - Use SQWRL to retrieve and instatiate Scenes (i.e. consecutive shots of the same cluster)
Scene Detection

Shot Clustering

- Create a 10-D vector composed of Audio And visual classes probability estimations
- Feed a MCL algorithm that uses Random Walks to create clusters
- Instantiate VSO:Cluster class in the VS Ontology
- Use SQWRL to retrieve and instatiate Scenes (i.e. consecutive shots of the same cluster)
Harmful Content Domain

Ontology Definition
Harmful Content Ontology Definition I

**Harmful Content Ontology**

- First attempt to conceptualize violence and pornography in an organized way
- Synthesis of psychologists’ violence definition and extensive investigation of movies depicting violent content
- Generic representation for use by psychologist, pedagogist, police
- Non violence classes implementation (scenery, dialogue, action) due to open world reasoning in OWL
Harmful Content Domain Ontology Definition II

Violence Ontology Defines:
- Complex Semantics of extensive violent acts
- Inter-relation of medium and low level semantics

Construction of violent actions hierarchy

*In a movie scene containing violence (e.g. torture, fight, war) a spectator can quickly grasp the form of violence (e.g. fighting without weapons), recognize a sequence of violent (e.g. punching, kicking), of generic (e.g. running, walking) and of consequence (e.g. falling, crawling, scream) actions.*

Association Mechanisms
- Medium level classes ↔ Inferred multimodal actions
- Low level concepts ↔ Represented in visual and audio ontology
Harmful Content Domain Ontology
Up to Now

- Preprocessing
- Visual analysis
- Audio analysis
- Scene Detection

- Video Structure Ontology
- Visual Ontology
- Audio Ontology
- Video Structure Ontology

- Harmful Content Ontology
- Inferencing Procedure
Up to Now

- Preprocessing
- Visual analysis
- Audio analysis
- Scene Detection

- Video Structure Ontology
- Visual Ontology
- Audio Ontology
- Video Structure Ontology

- Harmful Content Ontology
- Inferencing Procedure
Inferencing Procedure
Inferencing Procedure

Audio Classification

Visual Classification

Instantiation (Jena)

Instantiated Model

Video Structure Ontology

Visual Ontology

Audio Ontology

Violence Ontology

Video Ontology Classification (Pellet)

SWRL (1st Set) (Jess)

Ontology Classification (Pellet)

SWRL (2nd Set) (Jess)

Ontology Classification (Pellet)

Inference Engine

Non-Violence

Violence

Audio Ontology

Visual Ontology

Violence Ontology
Ontologies define the Terminological Box (TBox) of our knowledge base

The Instantiated model - ABox:
- Captures existing and extracted knowledge for the movie in question
- Derives directly from the **Segmentation, Audio, Visual and Scene analysis**
- Forms the basic facts in terms of individuals (shots, frames, events and objects)
- Includes low level numerical values and accuracy probability results
Instantiation

Property assertions: americanninja_01_avSegment_12
- hasModality AudioVisual
- hasPreviousSegment americanninja_01_avSegment_11
- hasNextSegment americanninja_01_avSegment_13

Data property assertions
- hasFightProbability 0.86153847
- hasEndingFrame 325
- hasVisualGunshotProbability 0.0
- hasVisualNoGunshotProbability 1.0
- hasSpeechProbability 0.98992807
- hasStartTimeMs 12000
- hasStartingFrame 300
- hasMusicProbability 0.006482982
- hasNormalActivityProbability 0.3312619
- hasNoActivityProbability 0.6687381
- hasSmoothSoundProbability 0.028125
- hasDuration 1000
- hasScreamProbability 0.003248511
- hasEndTimeMs 13000
- hasAudioGunshotProbability 0.016438356
- hasHighActivityProbability 0.0
- hasSharpSoundProbability 0.8

Property assertions: americanninja_01_frame_1736
- displayFace americanninja_01_visualObject_364
- displayFace americanninja_01_visualObject_383

Data property assertions
- hasFrameNumber 1736
- hasSkinPixelDiff 0.0
- hasObjectTemporalDiff 0.060363
- hasOrientationSddev 1.81153
- hasSkinPixelRatio 0.0
- hasOrientationMean 3.350859
- hasBiggestCCSize 0.0
- hasBiggestObjectSizeRatio 0.112617
- hasMaxLuminanceInterv 0.410156
- hasMaxLuminanceDiff 3.40625
- hasEkin 0.019766

Property assertions: americanninja_01_visualObject_383
- faceAppeareInFrame americanninja_01_frame_1750
- faceAppeareInFrame americanninja_01_frame_1739
- faceAppeareInFrame americanninja_01_frame_1750
- faceAppeareInFrame americanninja_01_frame_1735
- faceAppeareInFrame americanninja_01_frame_1753
The inference engine should take under consideration:

- Intra- and Cross-modality spatial, temporal or spatio-temporal relationships
- Importance of each modality for identifying a concept or semantic event
- Cross-modality synchronicity relationships (simultaneous semantic instances in different modalities)
- Uncertainty of extracted medium level semantics
- Support reasoning with partial, imprecise information
Inferencing Lifecycle I

Audio Classification

Visual Classification

Instantiation (Jena)

Instantiated Model

Video Structure Ontology

Audio Ontology

Visual Ontology

Violence Ontology

Inference Engine

Video Ontology Classification (Pellet)

SWRL (1st Set) (Jess)

Ontology Classification (Pellet)

SWRL (2nd Set) (Jess)

Ontology Classification (Pellet)

Audio Ontology

Violence Ontology

Non-Violence

Violence
## Inferencing Lifecycle II

### Steps:

1. Consistency check of the instantiated model and assertion of each individual’s initial class
2. First SWRL Set Rules Application
3. Consistency checking and classification services on the implied model
4. Second SWRL Set Rules Application
5. Consistency checking and classification services are applied to:
   - Infer violent and non violent segments (children to parents)
   - Extract extended semantics (parents to children)
Inferencing Lifecycle II

Steps:

1. Consistency check of the instantiated model and assertion of each individual’s initial class
2. First SWRL Set Rules Application
3. Consistency checking and classification services on the implied model
4. Second SWRL Set Rules Application
5. Consistency checking and classification services are applied to:
   - Infer violent and non violent segments (children to parents)
   - Extract extended semantics (parents to children)
### Inferencing Lifecycle II

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Consistency check of the instantiated model and assertion of each individual’s initial class</td>
</tr>
<tr>
<td>2</td>
<td>First SWRL Set Rules Application</td>
</tr>
<tr>
<td>3</td>
<td>Consistency checking and classification services on the implied model</td>
</tr>
<tr>
<td>4</td>
<td>Second SWRL Set Rules Application</td>
</tr>
<tr>
<td>5</td>
<td>Consistency checking and classification services are applied to:</td>
</tr>
<tr>
<td></td>
<td>- Infer violent and non violent segments (children to parents)</td>
</tr>
<tr>
<td></td>
<td>- Extract extended semantics (parents to children)</td>
</tr>
</tbody>
</table>
Inferencing Lifecycle II

Steps:

1. Consistency check of the instantiated model and assertion of each individual’s initial class

2. First SWRL Set Rules Application

3. Consistency checking and classification services on the implied model

4. Second SWRL Set Rules Application

5. Consistency checking and classification services are applied to:
   - Infer violent and non violent segments (children to parents)
   - Extract extended semantics (parents to children)
Inferencing Lifecycle II

Steps:

1. Consistency check of the instantiated model and assertion of each individual’s initial class
2. First SWRL Set Rules Application
3. Consistency checking and classification services on the implied model
4. Second SWRL Set Rules Application
5. Consistency checking and classification services are applied to:
   - Infer violent and non violent segments (children to parents)
   - Extract extended semantics (parents to children)
Inferencing Lifecycle II

Steps:

1. Consistency check of the instantiated model and assertion of each individual’s initial class
2. First SWRL Set Rules Application
3. Consistency checking and classification services on the implied model
4. Second SWRL Set Rules Application
5. Consistency checking and classification services are applied to:
   - Infer violent and non violent segments (children to parents)
   - Extract extended semantics (parents to children)
### Inferencing Lifecycle II

#### Steps:

1. Consistency check of the instantiated model and assertion of each individual’s initial class
2. First SWRL Set Rules Application
3. Consistency checking and classification services on the implied model
4. Second SWRL Set Rules Application
5. Consistency checking and classification services are applied to:
   - Infer violent and non violent segments (children to parents)
   - Extract extended semantics (parents to children)
Inferencing Lifecycle II

Steps:

1. Consistency check of the instantiated model and assertion of each individual’s initial class
2. First SWRL Set Rules Application
3. Consistency checking and classification services on the implied model
4. Second SWRL Set Rules Application
5. Consistency checking and classification services are applied to:
   - Infer violent and non violent segments (children to parents)
   - Extract extended semantics (parents to children)
### SWRL Example 1: Identifying a non violent (Activity, Action)

<table>
<thead>
<tr>
<th>Description</th>
<th>SWRL Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>If</strong></td>
<td></td>
</tr>
<tr>
<td>?avs is an Audio Visual Segment</td>
<td>VSO:AudioVisualSegment(?avs) ∧</td>
</tr>
<tr>
<td>?ms is an individual of audio class Music</td>
<td>MSO:Music(?ms) ∧</td>
</tr>
<tr>
<td>?ha is an individual of visual class High-Activity</td>
<td>MVisO:HighActivity(?ha) ∧</td>
</tr>
<tr>
<td>In ?avs music is detected</td>
<td>VSO:hasAudioEvent(?avs,?ms) ∧</td>
</tr>
<tr>
<td>In ?avs high activity is detected</td>
<td>VSO:hasVisualEvent(?avs,?ha) ∧</td>
</tr>
<tr>
<td><strong>Then</strong></td>
<td></td>
</tr>
<tr>
<td>?avs is an action segment</td>
<td>→ HCO:ActivityAction(?avs)</td>
</tr>
</tbody>
</table>
### SWRL Example 2: Identifying a violent (Fighting) segment

<table>
<thead>
<tr>
<th>Description</th>
<th>SWRL Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>If</td>
<td></td>
</tr>
<tr>
<td>?avs is an Audio Visual Segment</td>
<td>VSO:AudioVisualSegment(?avs) ∧</td>
</tr>
<tr>
<td>?f is an individual of audio class Fights</td>
<td>MSO:Fights(?g) ∧</td>
</tr>
<tr>
<td>?f is an individual of visual class HighActivity</td>
<td>MViso:HighActivity(?ha) ∧</td>
</tr>
<tr>
<td>In ?avs fights are detected</td>
<td>VSO:hasAudioEvent(?avs,?f) ∧</td>
</tr>
<tr>
<td>In ?avs high activity is detected</td>
<td>VSO:hasVisualEvent(?avs,?ha) ∧</td>
</tr>
<tr>
<td>Then</td>
<td></td>
</tr>
<tr>
<td>?avs is a fights segment</td>
<td>→ HCO:Fighting(?avs)</td>
</tr>
</tbody>
</table>
Inferring Higher Level Of Semantics

Identifying Person-On-Person-Fighting and Multiple-Person-Fighting

<table>
<thead>
<tr>
<th>HCO:Fighting Subclasses Definition</th>
<th>Necessary and Sufficient Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCO:PersonOnPersonFighting</td>
<td>HCO:displaysObjects some MVisO:Face ∧ HCO:displaysObjects exactly 2</td>
</tr>
<tr>
<td>HCO:MultiplePersonFighting</td>
<td>HCO:displaysObjects some MVisO:Face ∧ HCO:displaysObjects min 3</td>
</tr>
</tbody>
</table>
Implementation

- Matlab and the OpenCV library for audio/visual feature extraction and classification
- Protégé\textsuperscript{a} for the definition of ontologies and SWRL rules
- Pellet and Jess\textsuperscript{b} for ontology reasoning services and rules execution
- Jena semantic web framework\textsuperscript{c} for ontologies instantiation and synchronization of the knowledge generation lifecycle

---

\textsuperscript{a}http://protege.stanford.edu
\textsuperscript{b}http://www.jessrules.com/
\textsuperscript{c}http://jena.sourceforge.net/index.html
Dataset

- 50 videos have been extracted from 10 different movie films
- Total duration of the test data is 2.5 hours
- Per Shot/Segment manual annotation, using the Anvil video annotation research tool
- 19.4% of the data was of violent content
Results I

Measures

- **Precision** (i.e. the number of correctly detected violence segments, divided by the total number of detected violence segments)

- **Recall** (i.e. the number of correctly detected violence segments divided by the total number of true violence segments)

- **Mean Accuracy**

- $F_1 = \frac{2 \cdot P \cdot R}{P + R}$
## Results II

### Segment based Violence vs Non-violence Detection

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio-based</td>
<td>82.9%</td>
<td>38.9%</td>
<td>53%</td>
<td>61%</td>
</tr>
<tr>
<td>Visual-based</td>
<td>75.6%</td>
<td>34%</td>
<td>46.9%</td>
<td>54.8%</td>
</tr>
<tr>
<td><strong>Ontology-based</strong></td>
<td>91.2%</td>
<td>34.2%</td>
<td>50%</td>
<td>62.7%</td>
</tr>
</tbody>
</table>

### Ontological Segment based Multiclass Inference Measures

<table>
<thead>
<tr>
<th>Inference</th>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fights Inference</td>
<td>61.6%</td>
<td>68.2%</td>
<td>64.8%</td>
<td>64.9%</td>
</tr>
<tr>
<td>Screams Inference</td>
<td>41.4%</td>
<td>33.5%</td>
<td>37.1%</td>
<td>37.4%</td>
</tr>
<tr>
<td>Shots-Explosions Inference</td>
<td>63.3%</td>
<td>38.2%</td>
<td>47.6%</td>
<td>50.7%</td>
</tr>
</tbody>
</table>
## Results II

### Segment based Violence vs Non-violence Detection

#### Performance Measures

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Audio-based</strong></td>
<td>82.9%</td>
<td>38.9%</td>
<td>53%</td>
<td>61%</td>
</tr>
<tr>
<td><strong>Visual-based</strong></td>
<td>75.6%</td>
<td>34%</td>
<td>46.9%</td>
<td>54.8%</td>
</tr>
<tr>
<td><strong>Ontology-based</strong></td>
<td>91.2%</td>
<td>34.2%</td>
<td>50%</td>
<td>62.7%</td>
</tr>
</tbody>
</table>

### Ontological Segment based Multiclass Inference Measures

<table>
<thead>
<tr>
<th>Inference Type</th>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fights Inference</strong></td>
<td>61.6%</td>
<td>68.2%</td>
<td>64.8%</td>
<td>64.9%</td>
</tr>
<tr>
<td><strong>Screams Inference</strong></td>
<td>41.4%</td>
<td>33.5%</td>
<td>37.1%</td>
<td>37.4%</td>
</tr>
<tr>
<td><strong>Shots-Explosions Inference</strong></td>
<td>63.3%</td>
<td>38.2%</td>
<td>47.6%</td>
<td>50.7%</td>
</tr>
</tbody>
</table>
## Results III

### Ontological Shot based Binary and Multiclass Detection Performance Measures

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence Inference</td>
<td>61.07%</td>
<td>68.80%</td>
<td>64.7%</td>
<td>64.93%</td>
</tr>
<tr>
<td>Fights Inference</td>
<td>68.72%</td>
<td>89.9%</td>
<td>77.89%</td>
<td>79.31%</td>
</tr>
<tr>
<td>Screams Inference</td>
<td>25.0%</td>
<td>41.17%</td>
<td>31.10%</td>
<td>33.08%</td>
</tr>
<tr>
<td>Shots-Explosions Inference</td>
<td>89.39%</td>
<td>40.68%</td>
<td>55.91%</td>
<td>65.03%</td>
</tr>
</tbody>
</table>
Conclusions

Considering that

1. Extracted visual analysis clues are not at the desired level
2. Extracted audio and visual mid level clues are biased towards non-violence
3. Uncertain single modality results are treated as certain

We Conclude That

1. The attained results are really promising both for the binary and multiclass violence detection problem
2. The main advantage of using such an ontological approach still remains the higher level semantics extraction ability, using an unsupervised procedure and common sense reasoning
Consider that

1. Extracted visual analysis clues are not at the desired level
2. Extracted audio and visual mid level clues are biased towards non-violence
3. Uncertain single modality results are treated as certain

We Conclude That

1. The attained results are really promising both for the binary and multiclass violence detection problem
2. The main advantage of using such an ontological approach still remains the higher level semantics extraction ability, using an unsupervised procedure and common sense reasoning
Questions
An ontological approach to semantic video analysis for violence identification.
Best Paper Award in Multimedia Metadata Applications (M3A) Workshop.

Automatic movie content analysis - the MOCA project.

Gunshot detection in audio streams from movies by means of dynamic programming and bayesian networks.

Violence content classification using audio features.
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Existing Approaches</th>
<th>Ontological Methodology For Violence Identification</th>
<th>Conclusions</th>
</tr>
</thead>
</table>
Detecting pornographic video content by combining image features with motion information.


