CMU E-LAMP
Multimedia Event Recounting

TRECVID 2012

Nov 28, 2012

Outline

- MER 2012 Task
- CMU MER System
  - Feature Selection
  - Feature Integration
  - Recounting Presentation
- Submission
- Results and Analysis

MER 2012 Evaluation

- The system's recounting summarizations will be evaluated by a panel of judges
  - MER-to-Event: identify which event is represented by the recounting;
  - MER-to-Clip: match each of six MER outputs to the specific clip from which it was derived.

Presented Evidence (Features)

- Relationships
  - "Event-Relevant" Visual Concepts
  - "Event-Relevant" Audio Features
  - Co-occurrence (temporal) of Visual Concepts
- Observations
  - "Event-Specific" Visual Concepts
  - "Video-Specific" Visual Concepts
  - ASR Transcripts
  - Event-Specific Object Bank Results
  - Audio Concepts (Noisemes)

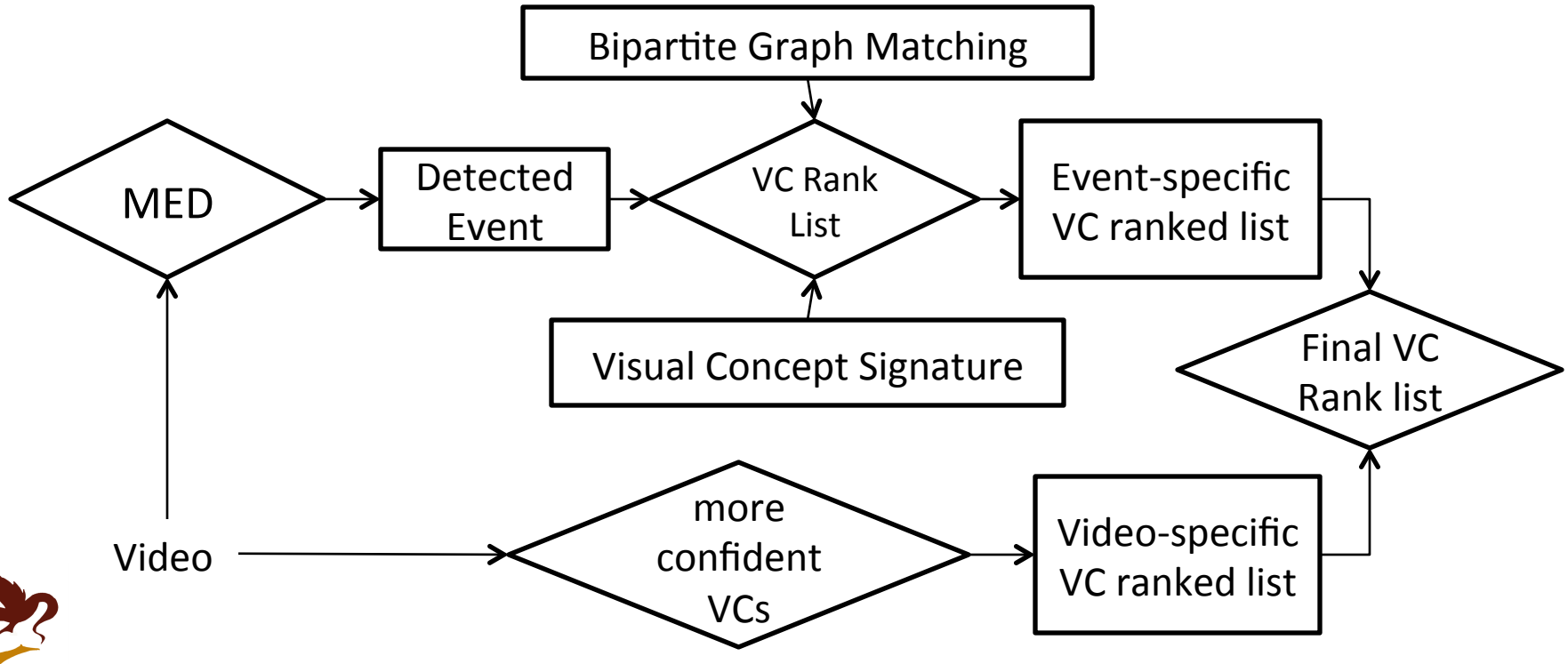
Presented Evidence (Features)

- Relationships
  - "Event-Relevant" Visual Concepts
  - "Event-Relevant" Audio Features
  - Co-occurrence (temporal) of Visual Concepts
- Observations
  - **"Event-Specific" Visual Concepts**
  - **"Video-Specific" Visual Concepts**
  - ASR Transcripts
  - Event-Specific Object Bank Results
  - Audio Concepts (Noisemes)

Ranking Visual Concepts

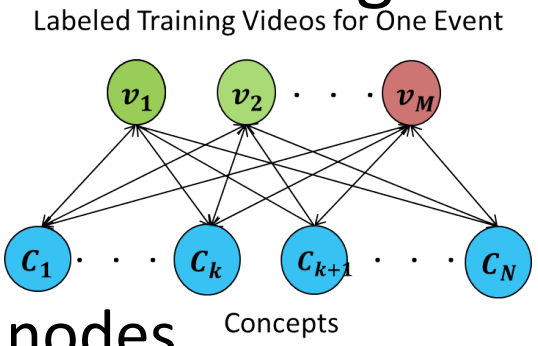
The goal is to list the top-n SIN visual concepts for a given video based on which

- One can predict the corresponding event → Relevance
- One can discriminate a given video from others → Uniqueness
- ObjectBank results were processed in the same way as the event-specific SIN VCs



Bipartite Graph Matching

- Explore pair-wise similarity between training videos and concepts
  - Video nodes
  - Concept nodes
  - Edges between video and concept nodes
  - Weight of each edge



- The concepts connecting to positive videos get higher scores
- The concepts connecting to negative videos get lower scores

Bipartite Graph Matching Semantic Classes to Events

- Example graph matching results

Events	Top 8 Semantic Classes
Flash mob gathering	Crowd, People_Marching, 3_Or_More_People, Demonstration_Or_Protest, Meeting, Cheering, Urban_Scenes, Walking
Getting a vehicle unstuck	Car, Snow, Motocycle, Outdoor, Landscape, Vehicle, Boad_Ship, Ground_Vehicles

ASR Transcripts

- Automatic speech recognition transcripts that capture “linguistic_audio” in the video.
  - E.g. “okay”, “this is where the crack was”, etc.
  - Segmentation based on power and time
- TF-IDF to calculate the saliency of each ASR results to the event kit.
  - Rank ASR transcripts according to relevance to event.
  - Include word posterior probabilities.

Audio Concepts (“Noisemes”)

- Noisemes the semantic audio concepts that indicate “non_linguistic_audio” information
  - E.g. “engine”, “music”, “animal” etc.
- Selection Strategy
  - Compute histogram of temporal distribution for video
  - Use bipartite graph matching to compute relevant sounds for each event
  - All the audio concepts are ranked for presentation based on their new scores

Presented Evidence (Features)

- Observations
  - "Event-Specific" Visual Concepts
  - "Video-Specific" Visual Concepts
  - ASR Transcripts
  - Event-Specific Object Bank Results
  - Audio Concepts (Noisemes)

Verification of UI and Selection

- "Are we presenting the right observations, or are we confusing users?"
  - Without actually performing the task
- Performed experiment to identify relevant and helpful observation groups
  - One-off diagnostic effort with small population
  - Iterative with MER developers for evaluation
- Could scale to larger setting

Imagine you need to describe the content (objects, actions, scenes, sounds, speech, and captions) of this video, to ident
"better" choice within each pair of features. This video is about "cleaning_an_appliance".

1. Which visual feature(s) describes the content better?

- Primate Eukaryotic_Organism Body_Parts Man_Made_Thing Amateur_Video
- Primate Apartments Eukaryotic_Organism Amateur_Video Man_Made_Thing

2. Which are the more important words that people said in this video?

- SHE DIDN'T HELLO THEY DON'T LIKE BUT THEY HAVE TO EITHER MOVE THAT I WOULD JUST
  GOD WE DO OUR BEST RATE AND HAVE YOU GOT THE WHOLE
- OKAY YES YEAH

3. Which caption(s) visible in this video are more helpful to identify the video's topic?

Results

- Overall quality of the features:
  - ASR is helpful
  - Object Bank and Audio Concepts are okay
  - Temporal information is important and useful, especially in distinguishing videos that are in the same event

CMU MER System Features

- SIN Visual Concepts
  - Event-Specific: >0.6, up to 4, ranked by importance
  - Video-Specific: >0.6, up to 6, ranked by importance
- Object Bank Visual Concepts
  - Threshold 0.6, up to 5, ranked by importance
- ASR Transcripts
  - Threshold 0.5, up to 5, ranked by importance
- Noisemes Audio Concepts
  - Threshold 0.1, up to 5, ranked by confidence
  - Importance = confidence * BGM_relevance

Relationships

The following visual observations were detected with high confidence in the video, and are relevant to the E026: Renovating_a_home event:	C = 0.19	I = 0.91
Room		
Construction_Worker		
Construction_Site		
Co-occurrence at 00:30: Roadway_Junction and Road	C = 0.89	I = 0.44
Roadway_Junction		
Road		
Co-occurrence at 00:35: Motorcycle and Traffic	C = 0.99	I = 0.36
Motorcycle		
Traffic		

Observations

scene	C = 0.52	I = 1.0
Room (whole video)		
person_s	C = 0.02	I = 0.9
Construction_Worker (whole video)		
object_s	C = 0.64	I = 0.8
Construction_Site (whole video)		
person_s	C = 1.0	I = 0.7
Attached_Body_Parts (whole video)		
scene	C = 0.35	I = 1.0
Daytime_Outdoor (keyframe anchored at 00:40)		
person_s	C = 0.4	I = 0.98
Military_Personnel (keyframe anchored at 00:04)		

Visual and Audio Relationships

Relationships

The following visual observations were detected with high confidence in the video, and are relevant to the Rock_climbing event:
Landscape
Gym
Mountain

The following audio observations were detected with high confidence in the video, and are relevant to the Rock_climbing event:
ALL RIGHT GOOD JOB
GOOD JOB YOU LOVE ME SO I SEE MYSELF
WHICH IS SO BOLD WAY OUT

Observations

scene
Landscape (whole video)

scene
Gym (whole video)

scene
Mountain (whole video)

Co-occurrence Relationship

Relationships

The following visual observations were detected with high confidence in the video, and a Rock_climbing event:

Landscape

Gym

Mountain

Co-occurrence at 03:23: Landscape and Mountain

Landscape

Mountain

Observations

scene

Landscape (whole video)

scene

Gym (whole video)

scene

Observations

scene	C = 0.95	I = 0.8
Kitchen (whole video)		
person_s	C = 0.65	I = 0.7
Hand (whole video)		
scene	C = 0.53	I = 1.0
Room (keyframe anchored at 00:26)		
scene	C = 0.57	I = 0.78
Kitchen (keyframe anchored at 01:07)		
person_s	C = 0.42	I = 0.67
Hand (keyframe anchored at 00:55)		
linguistic_audio	C = 0.97	I = 1.0
AND CLEAN THE HOUSE OF THEIR WAY WE DO THE BEST YOU SAID YOU KNOW HOW TO CLEAN ITSELF (00:06-00:11)		
linguistic_audio	C = 1.0	I = 0.91
CAMERAS ARE BIG BROTHER IS MY IMPORTANT DUTIES TELLS YOU HOW TO TAKE CARE EVERYONE (00:00-00:06)		
linguistic_audio	C = 0.65	I = 0.88
INSIDE THIS ANALOGY IS IF I SELL HIS NO VOTE HEAVILY SAYING THAT (01:24-01:29)		
linguistic_audio	C = 0.96	I = 0.86
AND THEN LET THEM INTERESTED SENOR IS WELL I'LL TELL YOU WHAT I DO WITH THE WORDS OF THE EASIEST WAY IS A VACUUM CLEANER (00:11-00:20)		
linguistic_audio	C = 0.74	I = 0.82
AND YOU BACK UNITS (00:21-00:23)		
object_s	C = 0.41	I = 1.0
stove		
object_s	C = 1.0	I = 0.9
television		

Observations

scene	C = 0.95	I = 0.8
Kitchen (whole video)		
person_s	C = 0.65	I = 0.7
Hand (whole video)		
scene	C = 0.53	I = 1.0
Room (keyframe anchored at 00:26)		
scene	C = 0.57	I = 0.78
Kitchen (keyframe anchored at 01:07)		
person_s	C = 0.42	I = 0.67
Hand (keyframe anchored at 00:55)		
linguistic_audio	C = 0.97	I = 1.0
AND CLEAN THE HOUSE OF THEIR WAY WE DO THE BEST YOU SAID YOU KNOW HOW TO CLEAN ITSELF (00:06-00:11)		
linguistic_audio	C = 1.0	I = 0.91
CAMERAS ARE BIG BROTHER IS MY IMPORTANT DUTIES TELLS YOU HOW TO TAKE CARE EVERYONE (00:00-00:06)		
linguistic_audio	C = 0.65	I = 0.88
INSIDE THIS ANALOGY IS IF I SELL HIS NO VOTE HEAVILY SAYING THAT (01:24-01:29)		
linguistic_audio	C = 0.96	I = 0.86
AND THEN LET THEM INTERESTED SENOR IS WELL I'LL TELL YOU WHAT I DO WITH THE WORDS OF THE EASIEST WAY IS A VACUUM CLEANER (00:11-00:20)		
linguistic_audio	C = 0.74	I = 0.82
AND YOU BACK UNITS (00:21-00:23)		
object_s	C = 0.41	I = 1.0
stove		
object_s	C = 1.0	I = 0.9
television		

Observations

scene
Kitchen (whole video)
person_s
Hand (whole video)
scene
Room (keyframe anchored at 00:26)
scene
Kitchen (keyframe anchored at 01:07)
person_s
Hand (keyframe anchored at 00:55)
linguistic_audio
AND CLEAN THE HOUSE OF THEIR WAY WE DO THE BEST YO

person_s

Hand (keyframe anchored at 00:55)

linguistic_audio

AND CLEAN THE HOUSE OF THEIR WAY WE DO THE BEST YOU SAID YOU

linguistic_audio

CAMERAS ARE BIG BROTHER IS MY IMPORTANT DUTIES TELLS YOU HOW

linguistic_audio

INSIDE THIS ANALOGY IS IF I SELL HIS NO VOTE HEAVILY SAYING THAT (01

linguistic_audio

AND THEN LET THEM INTERESTED SENOR IS WELL I'LL TELL YOU WHAT I
VACUUM CLEANER (00:11-00:20)

linguistic_audio

AND YOU BACK UNITS (00:21-00:23)

object_s

stove

object_s

television

rock

object_s

swing

object_s

gravel

object_s

coral

object_s

wall

non_linguistic_audio

music_sing (duration 00:55)

non_linguistic_audio

music (duration 00:28)

MER 2012 Submission

- Development Dataset for Dry Run (5*6 videos)
- Evaluation Dataset for Evaluation (5*6 videos)
- All “Positive” video clips from our MED system
  - E022 3878 cleaning_an_appliance
  - E026 3687 renovating_a_home
  - E027 3291 rock_climbing
  - E028 3154 town_hall_meeting
  - E030 3722 working_on_a_metal_crafts_project

  Total 17732

MER 2012 Results

	MER-to-Event	MER-to-Clip	Combined (0.4*E+0.6*C)
Average of submitted systems	68.75%	43.72%	0.54
CMU_ELamp-MER-System	85.56%	66.30%	**0.74**

- #2 of non-100% systems on MER-to-Event
- #1 on MER-to-Clip
- #1 in weighted ranking

Event Specific Results

Event Number	Event Name	MER-to-Event (Average)	Our Results
E022	Cleaning_an_appliance	70.02%	66.67%
E026	Renovating_a_home	67.70%	88.89%
E027	Rock_climbing	77.36%	96.30%
E028	Town_hall_meeting	86.75%	100.00%
E030	Working_on_a_metal_crafts_project	60.38%	75.93%

Video Specific Results

Event Number	Event Name	MER-to-Clip (Average)	Our Results
E022	Cleaning_an_appliance	53.67%	77.78%
E026	Renovating_a_home	40.99%	53.70%
E027	Rock_climbing	49.27%	64.81%
E028	Town_hall_meeting	36.23%	70.37%
E030	Working_on_a_metal_crafts_project	46.12%	64.81%

Future Work

- Integrate Optical Character Recognition
- Semantic recounting model for the event
  - Learn a better model to map the recountings of the video clips of an event to the event kit
- User study to benchmark selection and ranking strategies
- Perform more analysis on MER-to-Clip task

Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

Thank You