Pre-trained D-CNN Models for Detecting Complex Events in Unconstrained Videos

Joseph Robinson and Yun Fu

SPIE Commercial + Scientific Sensing and Imaging

Paper No. 9871-21

Synergetic Media Learning (SMILE) Laboratory Seminar

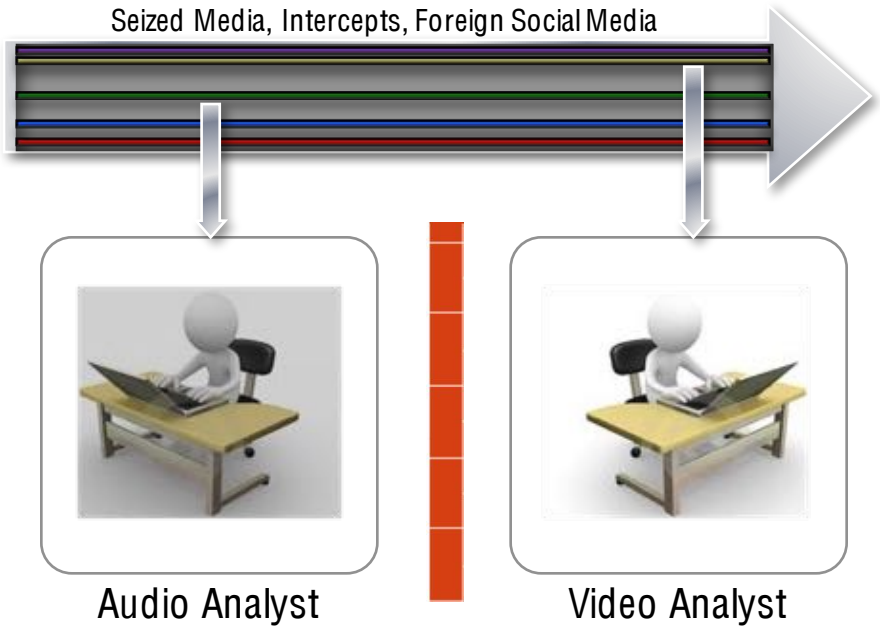
Department of Electrical and Computer Engineering

Northeastern University

18 April 2016

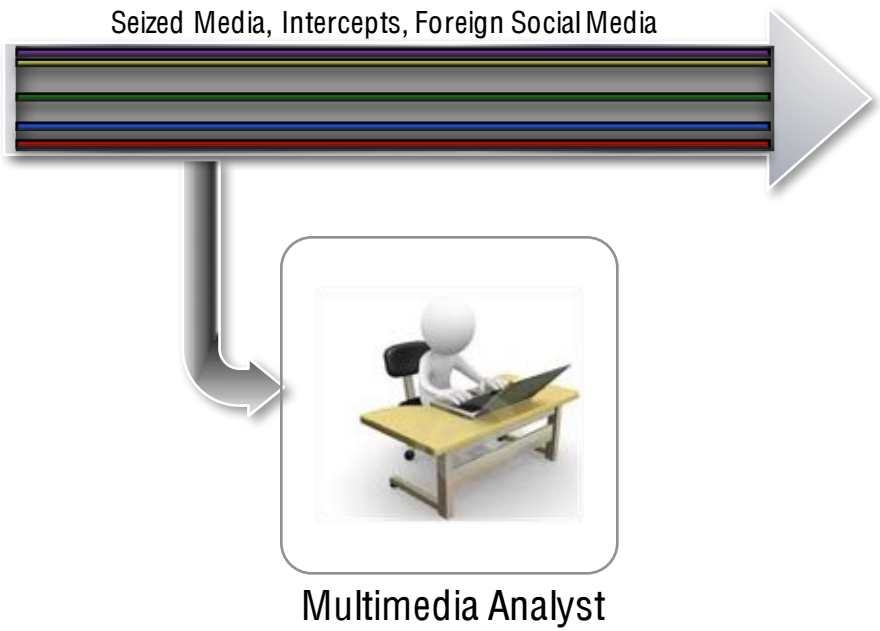
Multimedia Analyst Workflow

Without Cross-Media Exploitation



- USG currently relies on specialized analysts (e.g. audio speech analysts)
  - Stove-piped data and technical capabilities

With Cross-Media Exploitation



- Developing multimedia analysis capabilities
  - Enables cross-media analysis
  - Helps to address the deluge of data
  - Empowers non-specialized analysts

Multi-modal approaches exploit subtle connections which in turn enable deeper inference.

Multimedia Analyst Capabilities

Notional Capability

Search Query

Image Query:

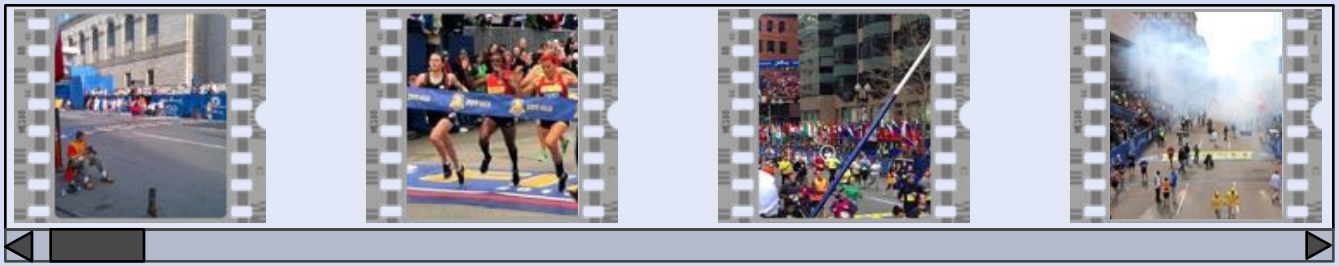
Text Query: Boston marathon

Constraint(s)

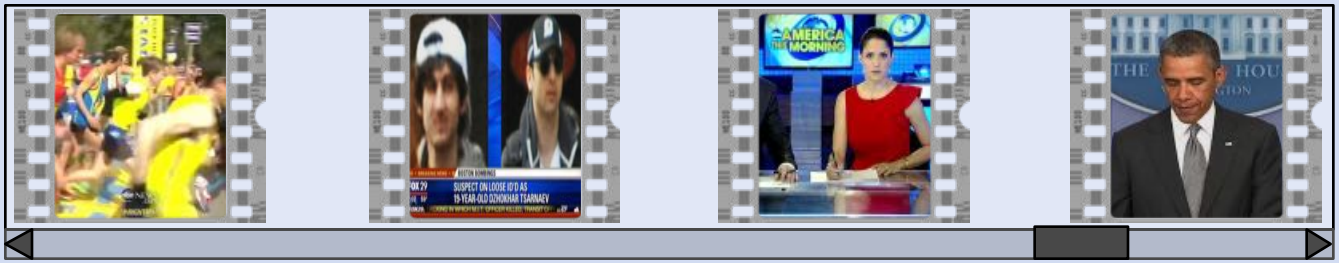
- outdoor scene
- contains people
- contains audio speech
- includes content

Attribute(s)

Scene Match ✓ ✗



Topic Match ✓ ✗



Video Archives

Growing Concerns

- YouTube has more than **1 billion** users
- Every day people watch hundreds of millions of hours on YouTube and generate billions of views
- **300 hours** of video are uploaded to YouTube every minute

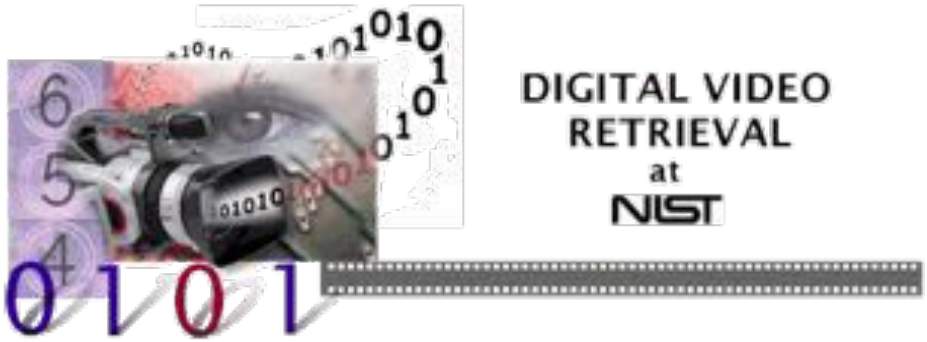
NEEDS

An efficient way to sort, search and retrieve this huge amount of video based on **CONTENT**

TRECVid

Overview

- **TRECVid**- Annual workshop series sponsored by NIST.
  - Extension of TREC (Text Retrieval Conference) series.
  - Initiated as *Video Track* in 2001; Gained independence in 2003.



Encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results.

What is TRECVid?

Involved Task Evaluations

1. Semantic Indexing (SIN)
   - Filtering, categorization, browsing, search according to concepts.
2. Interactive Surveillance Event Detection (SED)
   - Detection all occurrences of a given event in large amounts of surveillance video provided a textual description.
3. Instance Search (INS)
   - An archived video of a person, place, or thing of interest to you, known or unknown, the task is to find other videos of the same target, but not necessarily referencing the same modality (image or text).
4. **Multimedia Event Detection (MED)**
   - Search events provided via event kit (type, text, video exemplars).
   - Multimedia Event Recounting (MER): produce textual recounting that summarizes key evidence of the event.

TRECVid

MED Task

Given:

- An event kit which consists of an event name, definition, explication, video example.

Provide:

- A system that can search multimedia recordings for user defined events.

What is Video Event Detection?

- Complex activity taking place at a ***definite place*** and ***time***;
- Involves interaction between **people** with each other, *scenes,* and/or **objects**;
- Made up of various processes with inherited relationships bounded loosely, tightly, or both. Relationships are joined by ***temporal*** and ***semantic relationships***;
- **Directly observable**.

TRECVid

MED: Sample Event Kit

E-21

Attempting a board trick

TRECVid

MED: Sample Event Kit (continued)

E021

Event name: Attempting a bike trick

Definition: One or more people attempt to do a trick on a bicycle, motorcycle, or other type of motorized bike. To count as a bike for purposes of this event, the vehicle must have two wheels (excludes unicycles, ATVs, etc.).

Explication: Bikes are normally ridden with a person sitting down on seat and holding onto the handlebars and steering with their hands. Tricks consist of difficult ways of riding the bike, such as on one wheel, steering with feet or standing on the seat; or intentional motions made with the bike that are not simply slowing down/stopping the bike, propelling it forward, or steering the bike as it moves. Steering around obstacles or steering a bike off of a jump and landing on the ground are generally not considered tricks in and of themselves, however if the bike jump is set up so that the person is jumping over something, (e.g. jumping over people or vehicles or over a river), or if the person does a flip or other trick in the air, that would be considered a trick.

Evidential description:

Scene: **outside**, often in a skate park, parking lot or street

Objects/ People: **person riding a bike, bike**, ramps, helmet, conc

Activities: **riding bike on one wheel, standing on top of bik**

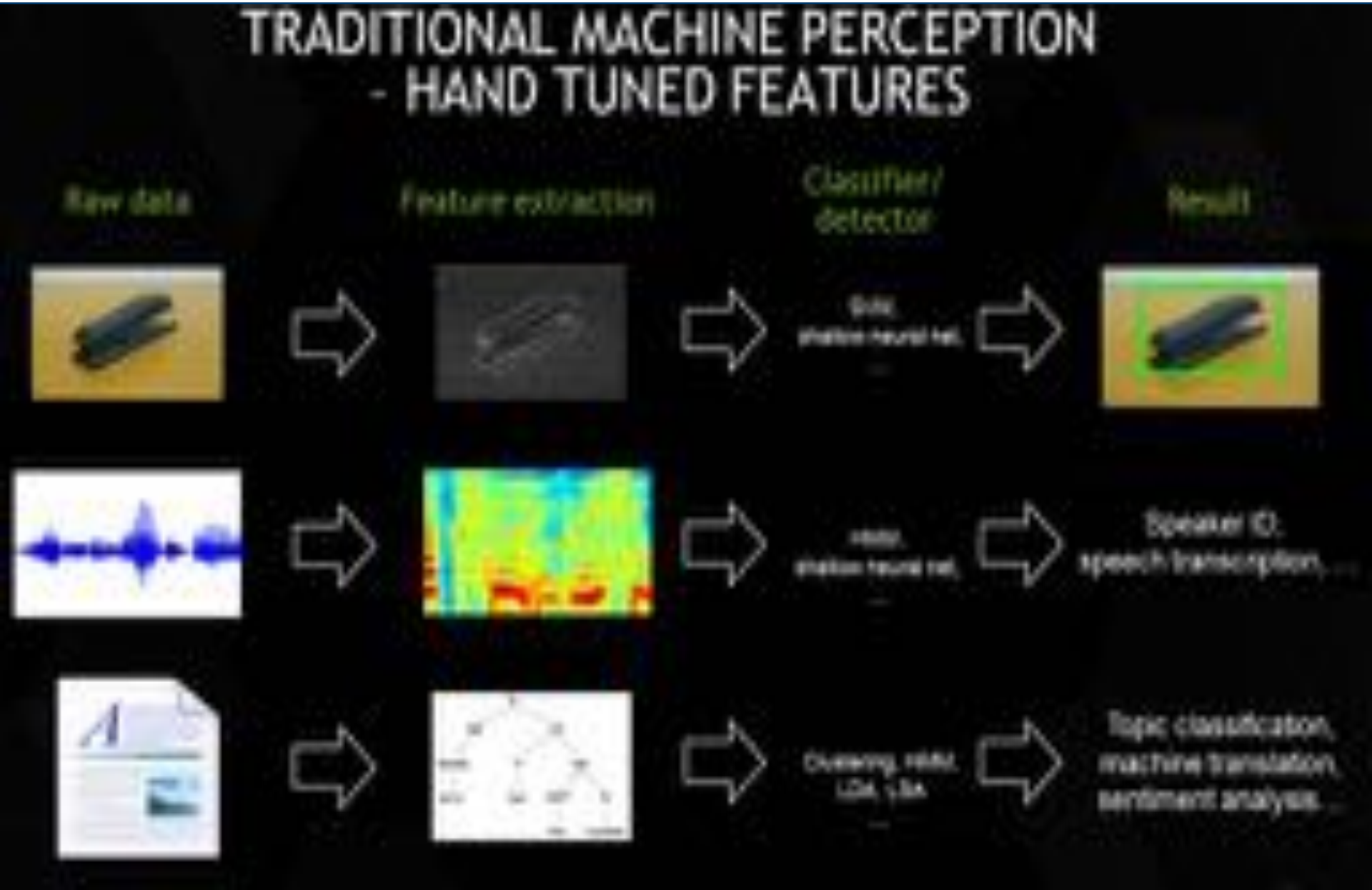
objects), spinning or flipping bike

Audio: sounds of bike hitting surface during the trick, audience cheering

Exemplar Videos

TRECVid

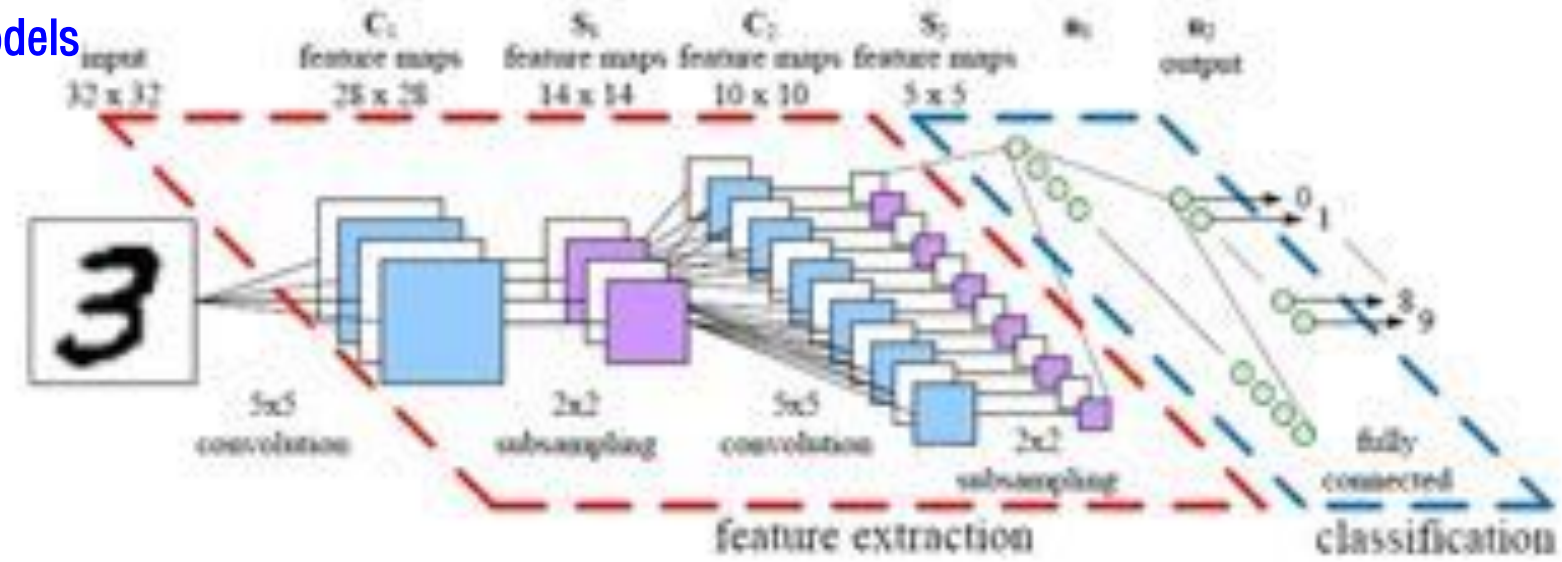
Approach



TRECVid

Approach

CNN Models



Pros

- Finds interesting features, i.e., no need for hand-crafted feature types.
- Current state-of-art in many vision-based tasks.
- Tools available online (e.g., Caffe).

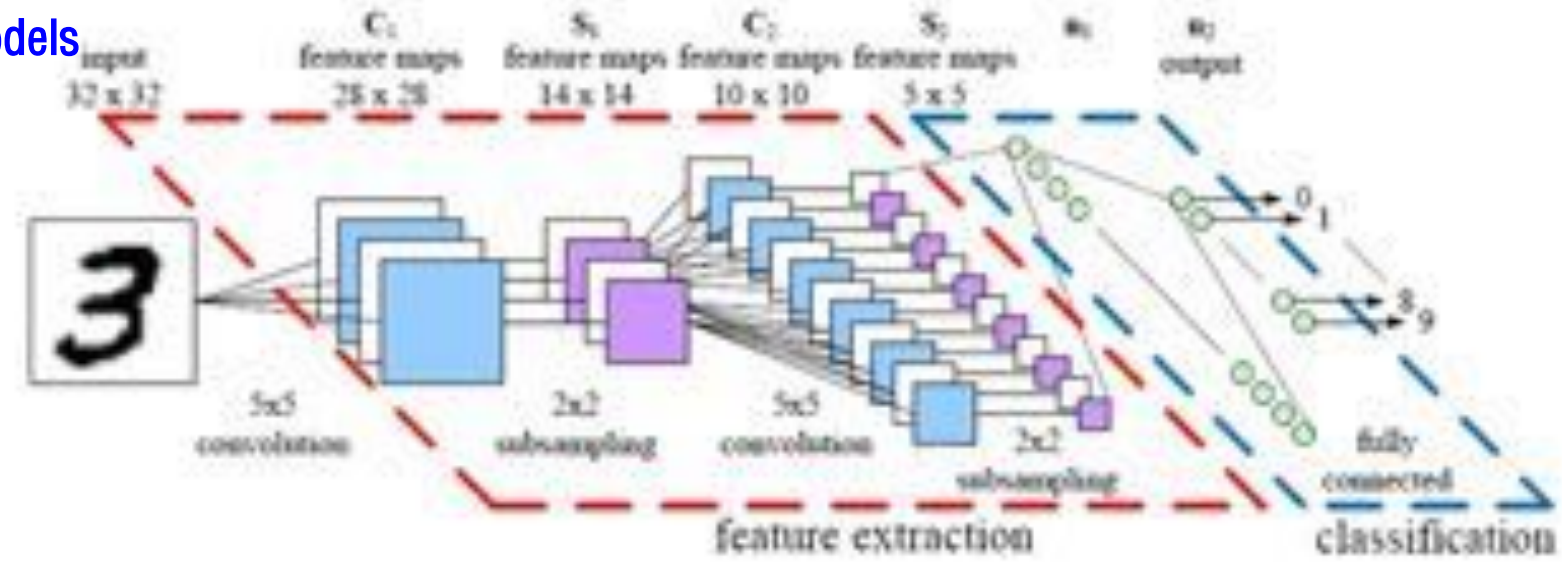
Cons

- Need lots of data.
- Takes time to train.

TRECVid

Approach

CNN Models



Pros

- Finds interesting features, i.e., no need for hand-crafted feature types.
- Current state-of-art in many vision-based tasks.
- Tools available online (e.g., Caffe).

Cons

- Need [redacted] data.
- Take[redacted] to train.

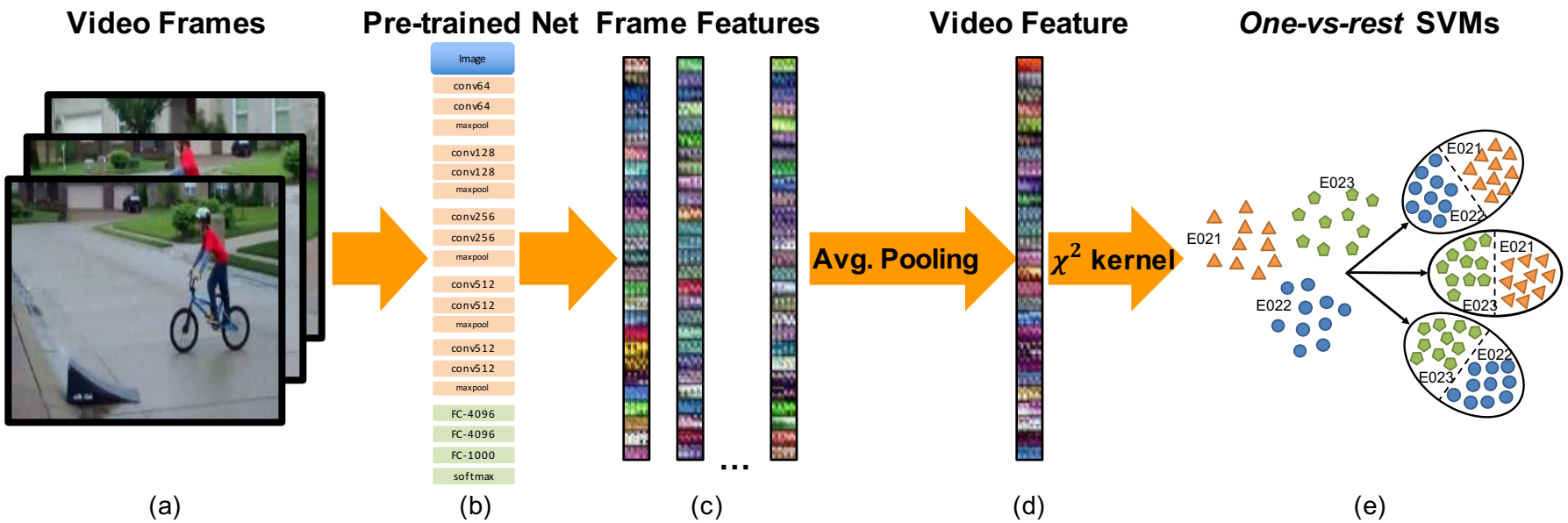
Pre-trained CNNs

Outline

- Motivation and Task Overview
- System Framework
  - Data
  - Feature Extraction
  - Encoding
  - Classification
- Results and Analysis
- Conclusions and Future Work
- Acknowledgements

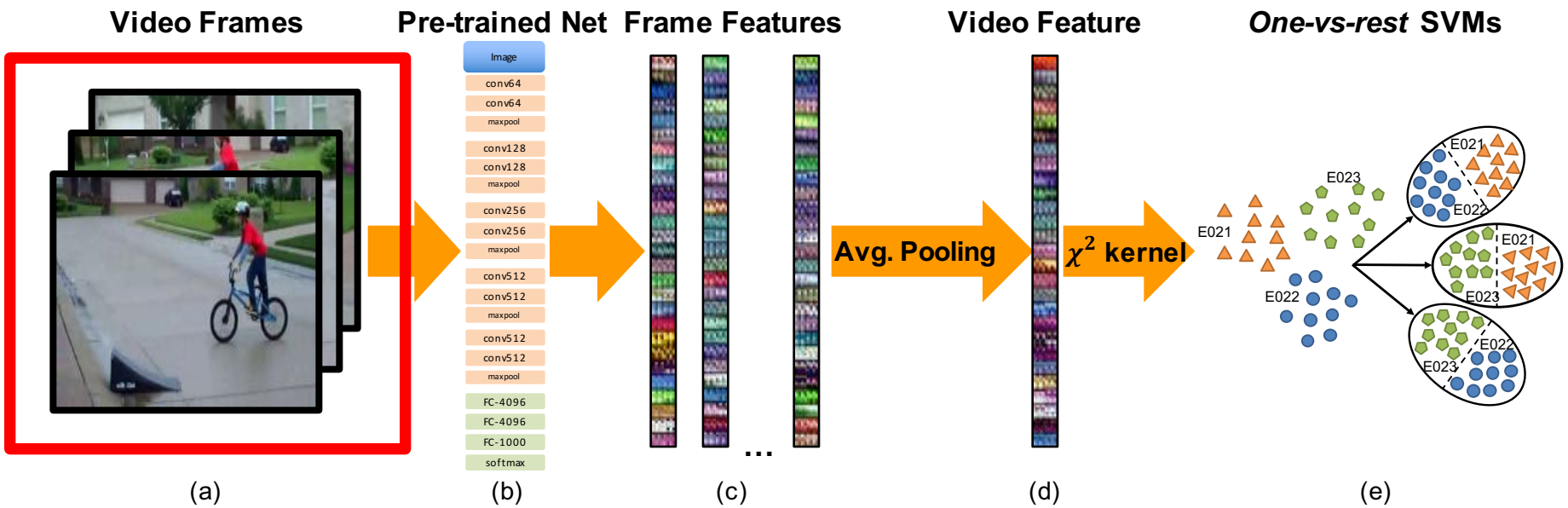
System Framework

Overview



System Framework

Sample Video Frames



FFMEG

- Sample 1/30 frames (approximately 1 fps).

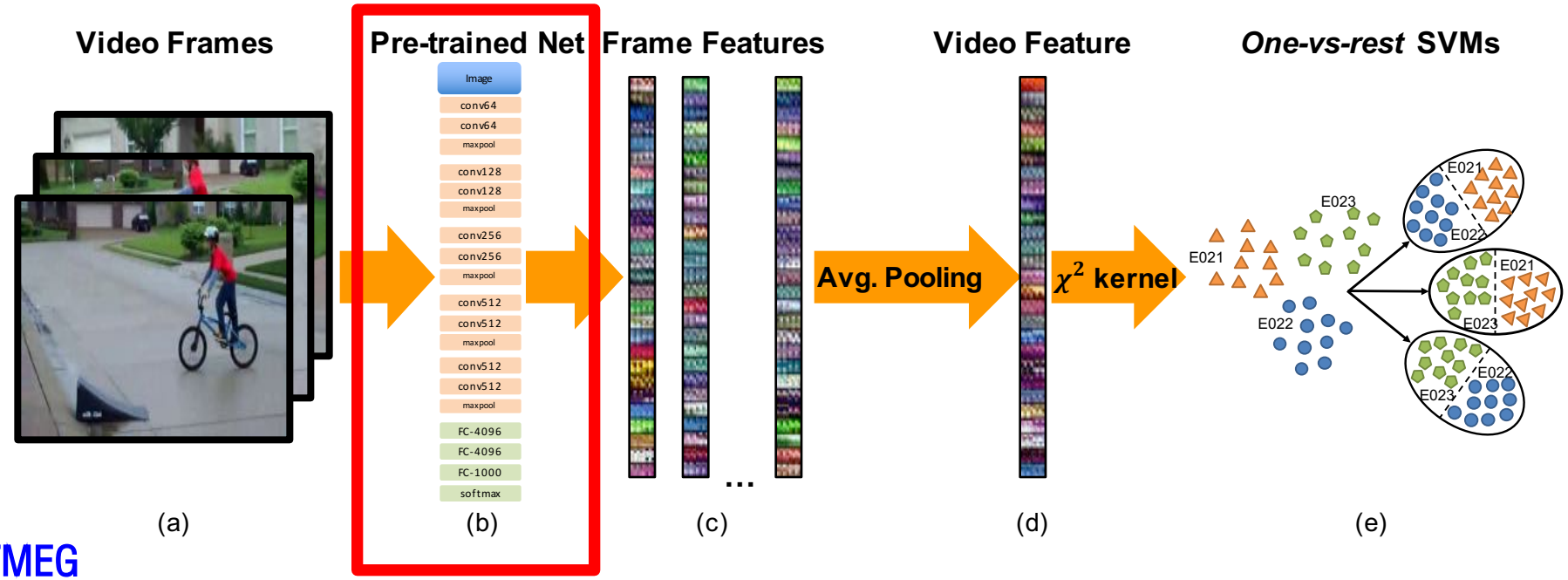
fps– frames per second

Outline

- Task Overview and Motivation
- System Framework
  - Data
  - Feature Extraction
  - Encoding
  - Classification
- Results and Analysis
- Conclusions and Future Work
- Acknowledgements

System Framework

Sample Video Frames



FFMEG

- Sample 1/30 frames (approximately 1 fps).

fps– frames per second

System Framework

Pre-trained CNN models

VGG-16

SPECS

- Very Deep Architecture
  - Trained weights using LFW
- Network Details
  - “Very small” convolution filters (i.e., 3x3)
  - Convolutional stride of 1
  - ReLu non-linearity
  - 3 Fully-Connected Layers

- Discretized into bins according to gradient [0°, 180°].
- Weighted (counted) according to magnitude of gradient.

Image
conv64
conv64
maxpool
conv128
conv128
maxpool
conv256
conv256
maxpool
conv512
conv512
maxpool
conv512
conv512
maxpool
FC-4096
FC-4096
FC-1000
softmax

System Framework

Feature Extraction

Places205

SPECS

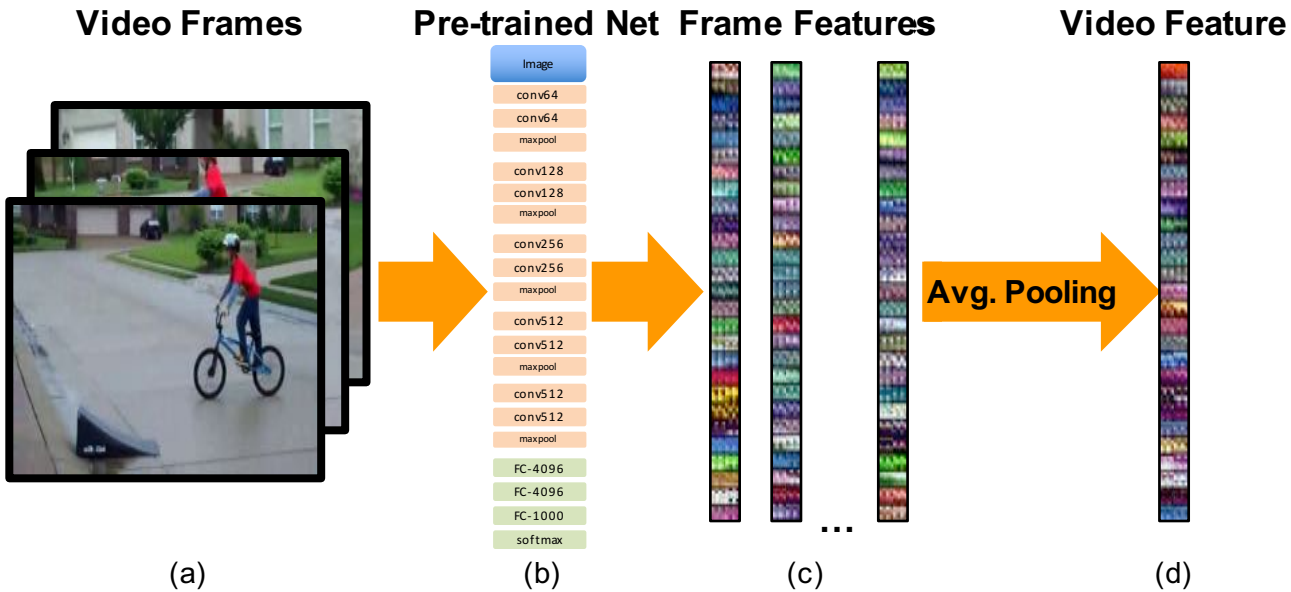
- Very Deep Architecture
  - Trained weights using LFW
- Network Details
  - “Very small” convolution filters (i.e., 3x3)
  - Convolutional stride of 1
  - ReLu non-linearity
  - 3 Fully-Connected Layers

- Discretized into bins according to gradient [0°, 180°].
- Weighted (counted) according to magnitude of gradient.

Image
conv64
conv64
maxpool
conv128
conv128
maxpool
conv256
conv256
maxpool
conv512
conv512
maxpool
conv512
conv512
maxpool
FC-4096
FC-4096
FC-1000
softmax

System Framework

Sample Video Frames

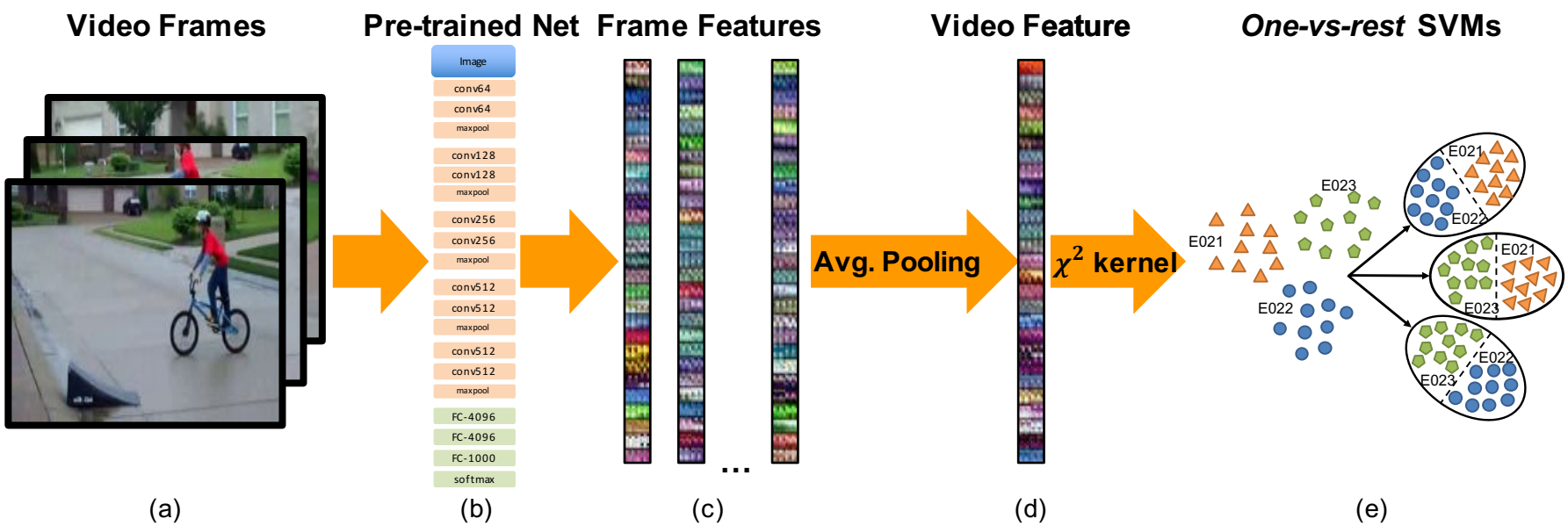


Average Pool Frame-level Features

$$\vec{x} = \frac{1}{N}\sum x_n$$

System Framework

Sample Video Frames



Outline

- Task Overview and Motivation
- System Framework
  - Data
  - Feature Extraction
  - Encoding
  - Classification
- **Results and Analysis**
- **Conclusions and Future Work**
- **Acknowledgements**

Results
Summarized

	Combinations	TP (%)
Run 1	Fc_7 features from VGG model (4,096D)	12.4
Run 2	Fc_8 features from VGG model (1,000D)	13.3
Run 3	Fc_8 features from Places205 model (205D)	8.4
Run 4	Fc_8 features from Hybrid-CNN model (1,186D)	11.7
Run 5	Fc_8 features from VGG (1,000D) and Places205 (205D) models; averaged SVM scores	11.8
Run 6	Fc_8 features from VGG and Places205 models concatenated (1,250D)	16.0

Results

Event-Specific

	Run-1	Run-2	Run-3	Run-4	Run-5	Run-6
E-21	17.2	**26.8**	5.8	21.4	17	19.4
E-22	2.2	2.4	2.4	1.4	2.5	**4**
E-23	**33.8**	26	14.1	24.8	26.9	33.6
E-24	0.4	1	0.7	0.8	1	**3.3**
E-25	0.2	0.3	**0.4**	0.3	0.3	0.3
E-26	7	8.8	8.7	5.2	**11.1**	10.5
E-27	**43.1**	32.5	18.1	29.7	25.4	36.8
E-28	14.3	10	11.1	9.6	11	**17.4**
E-29	25.4	16.9	22.7	12.9	19.2	**28.6**
E-30	8.3	9.4	3.8	5.3	8.8	**11.9**
E-31	23.7	**40.6**	8.5	37.2	18.8	25.6
E-32	**4.4**	3	3.5	2	3.6	3.6
E-33	6.3	10.8	2.6	4.5	9.8	**19.6**
E-34	6.1	4.7	2.2	2.2	3.7	**7.3**
E-35	18.9	22.5	18.3	24.9	23.3	**26**
E-36	2.8	3.3	2.5	3.9	3.2	**5.9**
E-37	5	7.6	6.4	7.3	6.7	**9.3**
E-38	2.2	2.2	4.6	3.7	3.5	**4.9**
E-39	14.6	25.2	25.7	27.8	30.5	**39.7**
E-40	11.5	11.2	6.5	8.5	9.6	**12.7**
Avg.	12.4	13.3	8.4	11.7	11.8	**16.0**

Results

Event-Specific

	Run-1	Run-2	Run-3	Run-4	Run-5	Run-6	
E-21	17.2	**26.8**	5.8	21.4	17	19.4	+5.4
E-22	2.2	2.4	2.4	1.4	2.5	**4**	
E-23	**33.8**	26	14.1	24.8	26.9	33.6	
E-24	0.4	1	0.7	0.8	1	**3.3**	
E-25	0.2	0.3	**0.4**	0.3	0.3	0.3	
E-26	7	8.8	8.7	5.2	**11.1**	10.5	
E-27	**43.1**	32.5	18.1	29.7	25.4	36.8	
E-28	14.3	10	11.1	9.6	11	**17.4**	-7.4
E-29	25.4	16.9	22.7	12.9	19.2	**28.6**	
E-30	8.3	9.4	3.8	5.3	8.8	**11.9**	
E-31	23.7	**40.6**	8.5	37.2	18.8	25.6	+15.0
E-32	**4.4**	3	3.5	2	3.6	3.6	
E-33	6.3	10.8	2.6	4.5	9.8	**19.6**	-8.8
E-34	6.1	4.7	2.2	2.2	3.7	**7.3**	
E-35	18.9	22.5	18.3	24.9	23.3	**26**	
E-36	2.8	3.3	2.5	3.9	3.2	**5.9**	
E-37	5	7.6	6.4	7.3	6.7	**9.3**	
E-38	2.2	2.2	4.6	3.7	3.5	**4.9**	
E-39	14.6	25.2	25.7	27.8	30.5	**39.7**	
E-40	11.5	11.2	6.5	8.5	9.6	**12.7**	
Avg.	12.4	13.3	8.4	11.7	11.8	**16.0**	

Analysis

Extreme Cases

		Top Activation	2nd Activation	3rd Activation
+5.4	E021	**Unicycle**  Vehicle with 1 wheel driven by pedals	**Tricycle**  Vehicle with 3 wheels driven by pedals	**All-Terrain Bike**  Bike w sturdy frame + fat tires; designed for mountainous country
+15.0	E031	**Bee House**  Shed containing a number of beehives	**Honeycomb**  Tiny hex-cells of beeswax used by bees to store honey + larvae	**Poncho**  Blanket-like cloak with a hole centered for head
-7.4	E028	**Theater Curtain**  Cloth, stage front; opens to start + closes to break/end	**Home Theater**  TV + video equipment to provide theater movie experience at home	**Volleyball**  An inflated ball used in playing volleyball
-8.8	E033	**All-Terrain Bike**  Bike w sturdy frame + fat tires; designed for mountainous country	**Disc Brake**  Brakes that applies friction to spinning disk via brake pads	**Tandem Bike**  Bike with 2 sets of pedals and 2 seats

Analysis

Extreme Cases: TP

	Top Activation	2nd Activation	3rd Activation
E021	**Unicycle**  Vehicle with 1 wheel driven by pedals	**Tricycle**  Vehicle with 3 wheels driven by pedals	**All-Terrain Bike**  Bike w sturdy frame + fat tires; designed for mountainous country
E031	**Bee House**  Shed containing a number of beehives	**Honeycomb**  Tiny hex-cells of beeswax used by bees to store honey + larvae	**Poncho**  Blanket-like cloak with a hole centered for head
E028	**Theater Curtain**  Cloth, stage front; opens to start + closes to break/end	**Home Theater**  TV + video equipment to provide theater movie experience at home	**Volleyball**  An inflated ball used in playing volleyball
E033	**All-Terrain Bike**  Bike w sturdy frame + fat tires; designed for mountainous country	**Disc Brake**  Brakes that applies friction to spinning disk via brake pads	**Tandem Bike**  Bike with 2 sets of pedals and 2 seats

TP– True Positive

Analysis

Extreme Cases: TP

	Top Activation	2nd Activation	3rd Activation
E021	**Unicycle**  Vehicle with 1 wheel driven by pedals	**Tricycle**  Vehicle with 3 wheels driven by pedals	**All-Terrain Bike**  Bike w sturdy frame + fat tires; designed for mountainous country
E031	**Bee House**  Shed containing a number of beehives	**Honeycomb**  Tiny hex-cells of beeswax used by bees to store honey + larvae	**Poncho**  Blanket-like cloak with a hole centered for head

Analysis

E-21

Analysis

E-31

Analysis

Extreme Cases: FP

	Top Activation	2nd Activation	3rd Activation
E021	**Unicycle**  Vehicle with 1 wheel driven by pedals	**Tricycle**  Vehicle with 3 wheels driven by pedals	**All-Terrain Bike**  Bike w sturdy frame + fat tires; designed for mountainous country
E031	**Bee House**  Shed containing a number of beehives	**Honeycomb**  Tiny hex-cells of beeswax used by bees to store honey + larvae	**Poncho**  Blanket-like cloak with a hole centered for head
E028	**Theater Curtain**  Cloth, stage front; opens to start + closes to break/end	**Home Theater**  TV + video equipment to provide theater movie experience at home	**Volleyball**  An inflated ball used in playing volleyball
E033	**All-Terrain Bike**  Bike w sturdy frame + fat tires; designed for mountainous country	**Disc Brake**  Brakes that applies friction to spinning disk via brake pads	**Tandem Bike**  Bike with 2 sets of pedals and 2 seats

FP– False Positive

Analysis

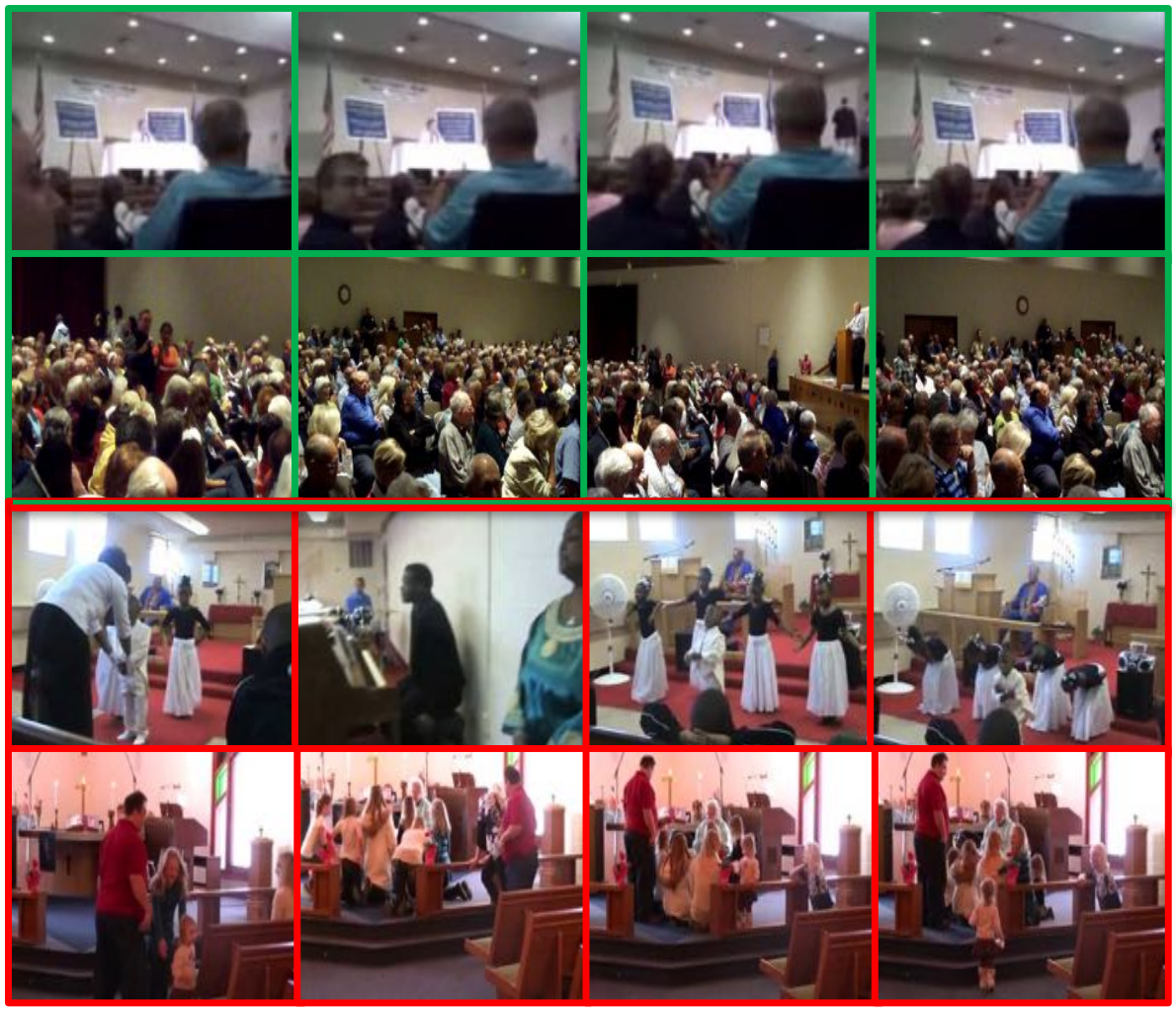
Extreme Cases: FP

	Top Activation	2nd Activation	3rd Activation
E028	**Theater Curtain**  Cloth, stage front; opens to start + closes to break/end	**Home Theater**  TV + video equipment to provide theater movie experience at home	**Volleyball**  An inflated ball used in playing volleyball
E033	**All-Terrain Bike**  Bike w sturdy frame + fat tires; designed for mountainous country	**Disc Brake**  Brakes that applies friction to spinning disk via brake pads	**Tandem Bike**  Bike with 2 sets of pedals and 2 seats

Analysis

Extreme Cases: FP

E-28



FP– False Positive

Analysis

Extreme Cases: FP

E-33

FP– False Positive

Outline

- Task Overview and Motivation
- System Framework
  - Data
  - Feature Extraction
  - Encoding
  - Classification
- Results and Analysis
- **Conclusions and Future Work**
- **Acknowledgements**

Multimedia Analyst Capabilities

Notional Capability

Search Query

Image Query:

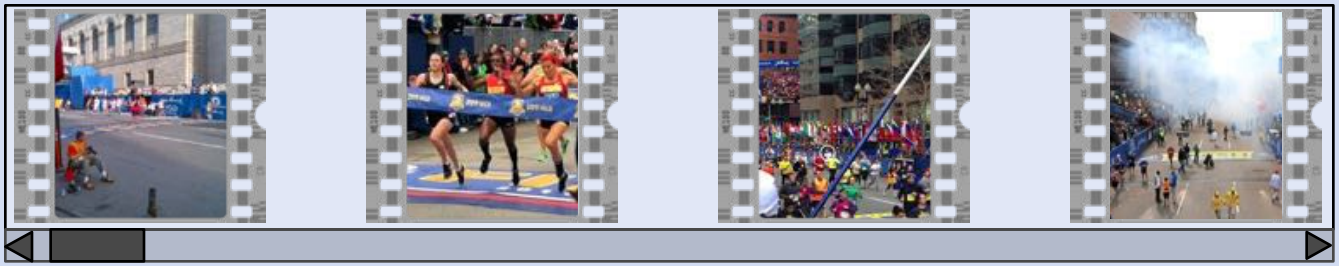
Text Query: Boston marathon

Constraint(s)

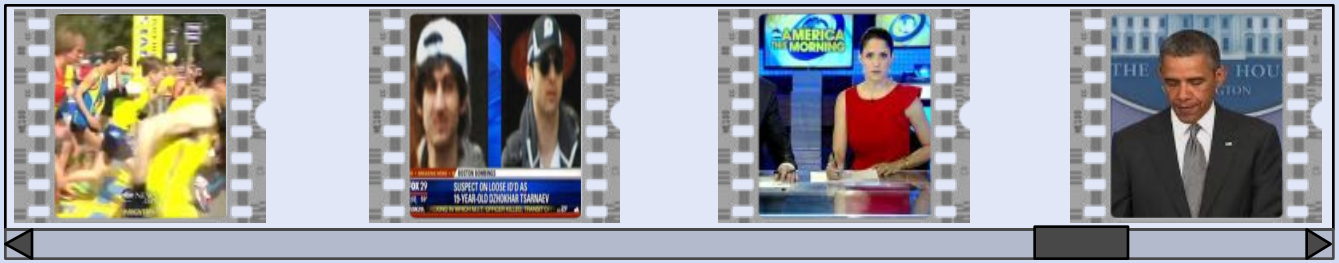
- outdoor scene
- contains people
- contains audio speech
- includes content

Attribute(s)

Scene Match ✓ ✗



Topic Match ✓ ✗



Multimedia Analyst Capabilities

Notional Capability

Image Query:

Text Query: Boston marathon

Audio Query:

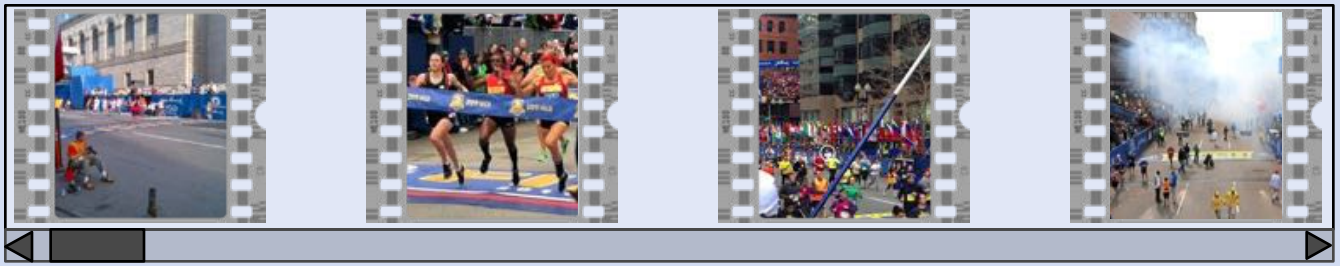
Constraint(s)

- outdoor scene
- contains people
- contains audio speech
- includes content

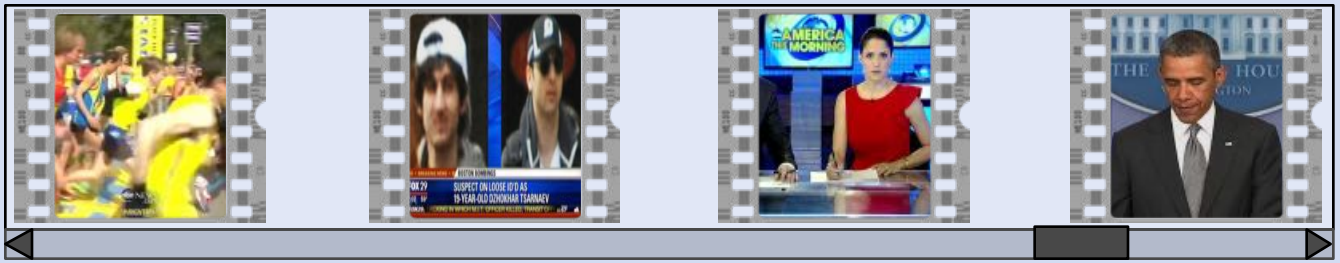
[X] Videos [ ] Records [X] Speech

Attribute(s)

Scene Match ✓ ✗



Topic Match ✓ ✗



Conclusion

End-to-end System

- Built system to detect events in videos w deep CNNs as *off-the-shelf* feature extractors.
- System SW is simple + modular– will be available project page soon!

Analysis

- Provided some insight about performance on difference CNN models for different events.
  - Presence of objects typically discriminates better than that off scene-types.
  - Fusion of object and scene based CNNs works better than Hybrid-CNN model.
  - Great promise for training deep CNN for object classes that are *event specific.*

Baseline

- Used system for TRECVid debut.
- Baseline from which information from other modalities can be incorporated in.
- Also, temporal information and co-occurrence dependencies will be added in the future.

References

[1] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. *Caffe: Convolutional architecture for fast feature embedding.* arXiv preprint arXiv:1408.5093 (2014).

[2] Vedaldi, A., & Fulkerson, B. (200 8). *VLFeat: An Open and Portable Library of Computer Vision Algorithms.* In ACM MM (2010).

[3] Wang ,L., Guo ,S., Huang ,W., and Qiao,Y., *Places205-vgg net models for scene recognition. CoRR* abs/1508.01667 (2015).

[4] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., *Imagenet: A large-scale hierarchical image database.* In Proceedings of IEEE CVPR (2009).

[5] Zhou,B., Lapedriza,A., Xiao,J., Torralba,A., & Oliva,A. *Learning deep features for scene recognition using places database.* NIPS (2014).

[6] Simonyan, K. and Zisserman, A., *Very deep convolutional networks for large-scale image recognition.* CoRR abs/1409.1556 (2014).

[7] Robinson, J. P., Scott, E., and Fu, Y., *NEU MIT-LL @ TRECVid 2015: Multimedia Event Detection by Deep Feature Learning.* In Proceedings of TRECVID 2015 (2015).

[8] Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A. F., Quenot, G., and Ordelman, R., *Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics.* In Proceedings of TRECVID (2015).

[9] Hsu, C.W. & Lin, C.J. *A comparison of methods for multiclass support vector machines*. IEEE Transactions on Neural Networks (2002).

[10] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L., *Imagenet large scale visual recognition challenge.* IJCV (April 2015).

[11] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., *Going deeper with convolutions*. In Proceedings of IEEE CVPR (2015).

Thank you!