ACTIVE: Activity Concept Transitions in Video Event Classification

Chen Sun and Ram Nevatia {chensun|nevatia}@usc.edu

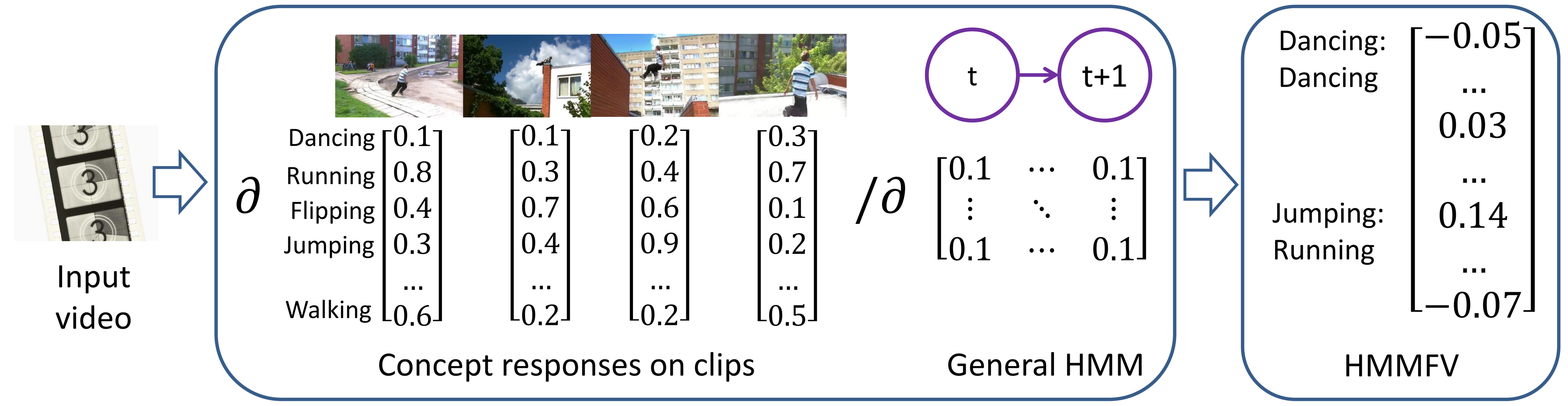
Institute for Robotics and Intelligent Systems, Viterbi School of Engineering, University of Southern California

Introduction

- **Goal**
  - Classify high level events from unconstrained web videos
- **Challenges**
  - Complex human-object interactions
  - Diverse video quality (e.g. YouTube)
  - Large scale dataset
- **Motivation**
  - Encoding activity concept transitions

	kiss	walk	dance
kiss	0.08	-0.02	0.06
walk	-0.07	-0.03	0.01
dance	-0.02	0.05	0.14

Hidden Markov Model Fisher Vector (HMMFV)



- **Video Representation:** A sequence of activity concept responses
- **Fisher Kernel**: Partial derivatives of log-likelihood function
- **HMMFV**: Partial derivatives about HMM transition parameters

$$\frac{\partial}{\partial \tau_{i|j}} \log P(X|\theta,\tau) = \sum_t \left[\frac{\xi_t(i,j)}{\tau_{i|j}} - \gamma_{t-1}(j) \right]$$

$$FV(i,j) \sim \sum_t \alpha_{t-1}(j) \left[\theta_{x_t|i} \beta_t(i) - \beta_{t-1}(j) \right]$$

$$\xi_t(i,j) = P(s_t = i, s_{t-1} = j | X, \theta, \tau)$$

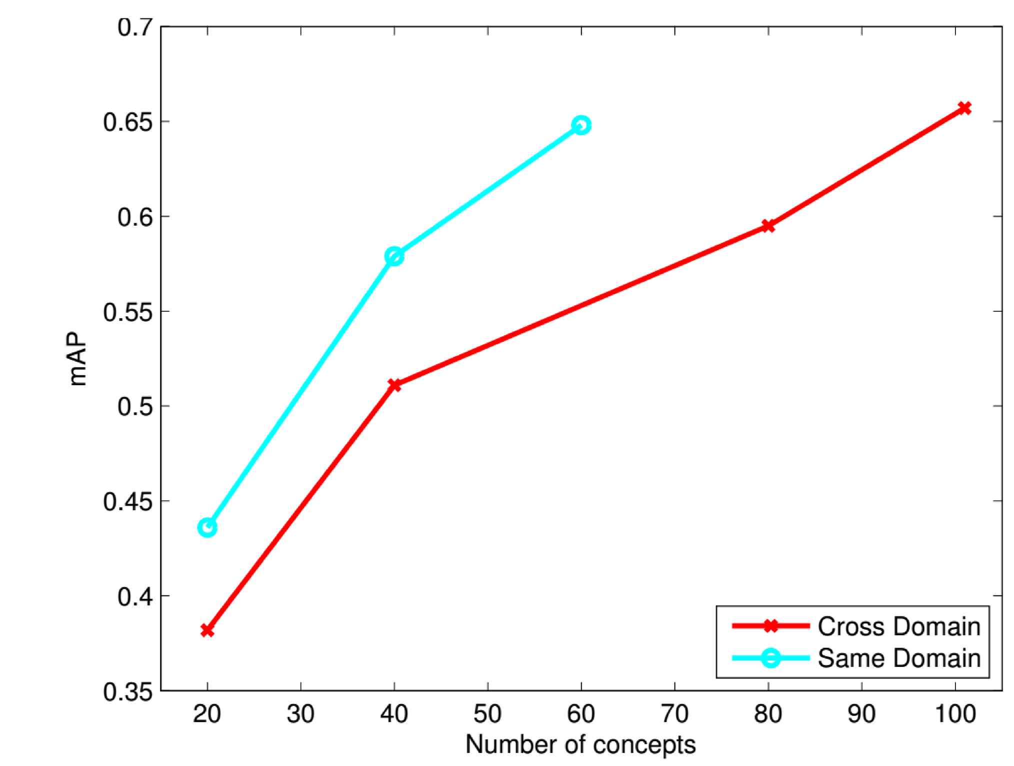
$$\gamma_{t-1}(j) = P(s_{t-1} = j | X, \theta, \tau)$$

$$\alpha_t(i) = P(x_1, ..., x_t, s_t = i | \theta, \tau)$$

$$\beta_t(i) = P(x_{t+1}, ..., x_n | s_t = i, \theta, \tau)$$

Experimental Setup

- **Dataset**
  - TRECVID MED 2011 Event Kit
  - 70% for training, 30% for testing
- **Setup**
  - Gaussian kernel SVM, 5-fold cross validation
- **Activity Concepts**
  - Same domain (Event Kit annotations [1])
  - Cross domain (UCF 101)



Top Activity Concept Transitions Visualization

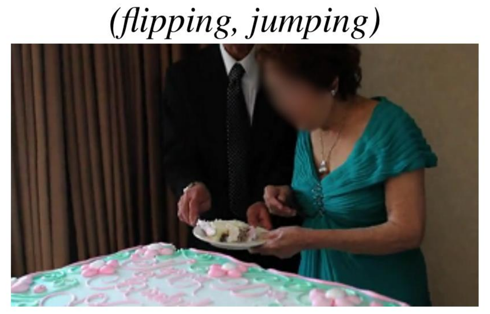
attempting a board trick
(jumping, jumping)
(sliding, jumping)
(flipping, jumping)

feeding an animal
(reeling, reeling)
(hands visible, hands visible)
(walking, animal eating)

landing a fish
(reeling, reeling)
(casting, reeling)
(hands visible, reeling)

wedding ceremony
(kissing, kissing)
(pouring, kissing)
(hugging, kissing)

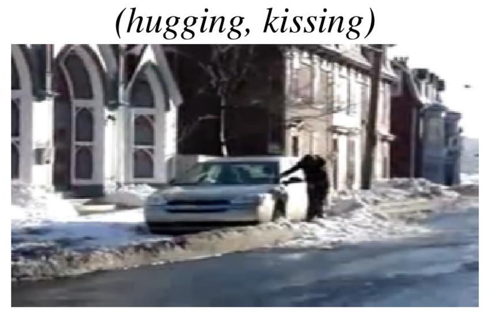
woodworking project
(hands visible, hands visible)
(carving, carving)
(carving, hands visible)



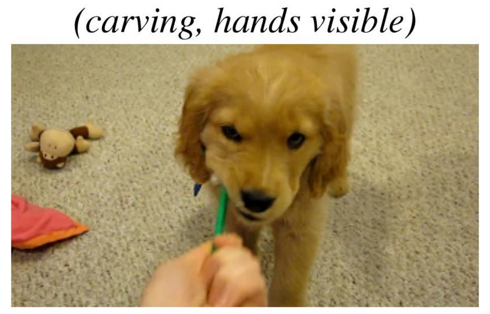
birthday party
(singing, singing)
(singing, jumping)
(clapping, singing)

changing a vehicle tire
(turning wrench, turning wrench)
(hands visible, turning wrench)
(fitting bolts, turning wrench)

flash mob gathering
(marching, marching)
(dancing, marching)
(dancing, dancing)



getting a vehicle unstuck
(steering, steering)
(vehicle moving, steering)
(vehicle moving, vehicle moving)



grooming an animal
(washing, washing)
(hands visible, washing)
(hands visible, hands visible)

making a sandwich
(hands visible, hands visible)
(spreading cream, spreading cream)
(spreading cream, hands visible)

parade
(marching, marching)
(dancing, marching)
(walking, marching)

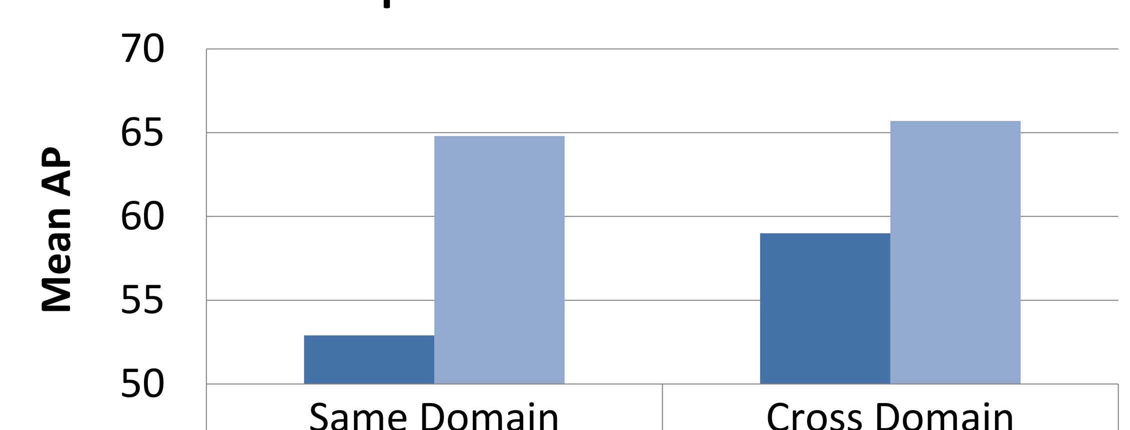
parkour
(jumping, jumping)
(flipping, jumping)
(dancing, jumping)

repairing an appliance
(hands visible, hands visible)
(turning wrench, hands visible)
(pointing, hands visible)

sewing project
(sewing, sewing)
(hands visible, sewing)
(eating, sewing)

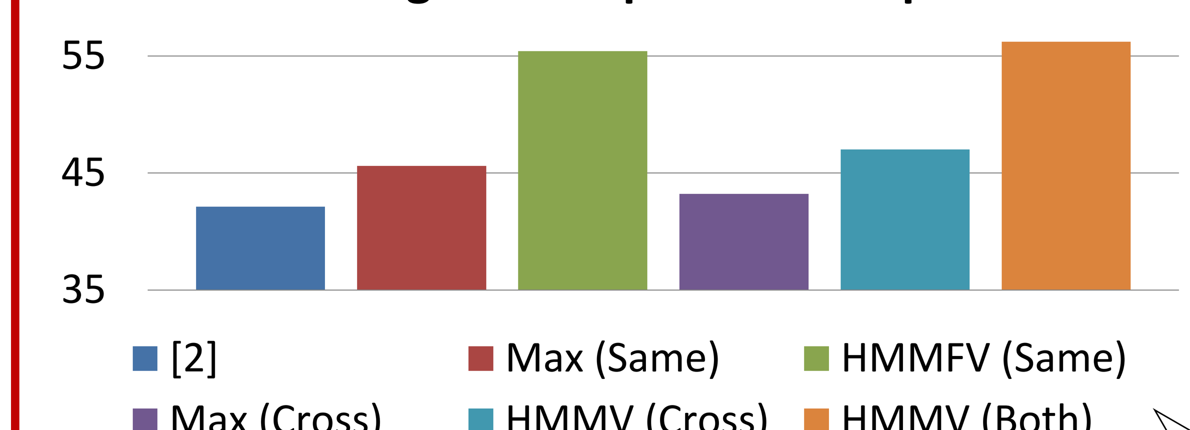
Quantitative results

Comparison with baseline



	Same Domain	Cross Domain
Max	52.9	59
HMMFV	64.8	65.7

Training with 10 positive samples



Comparison with state-of-the-art

Event	Joint [1]	LL [2]	HMMFV	HMMFV+[2]
Board trick	75.7	81.3	**88.5**	88.2
Feed animal	**56.5**	36.3	46.8	46.1
Wedding	72.2	74.3	**79.6**	78.9
Land fish	67.5	**82.9**	77.1	81.1
Woodworking	**65.3**	49.6	60.4	62.3
Birthday party	78.2	73.6	81.1	**81.4**
Change tire	47.7	**54.1**	48.2	51.8
Flash mob	**91.9**	86.8	87.3	87.7
Vehicle unstuck	69.1	76.9	75.6	**77.2**
Groom animal	51.0	57.9	61.7	**63.4**
Make sandwich	41.9	51.5	47.6	**52.4**
Parade	72.4	72.0	77.0	**77.0**
Parkour	66.4	79.2	88.6	**89.0**
Repair appliance	**78.2**	66.1	61.9	63.4
Sewing	57.5	60.8	57.5	**62.1**
Mean AP	66.1	66.9	69.3	**70.8**

[1] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In ECCV, 2012.
[2] C. Sun and R. Nevatia. Large-scale Web Video Event Classification by use of Fisher Vectors. In WACV, 2013

Conclusion

- Coding temporal transitions of activities by HMMFV improves performance
- Activity concepts coded by HMMFV is desirable with limited training samples
- May be useful for video event recounting (description)