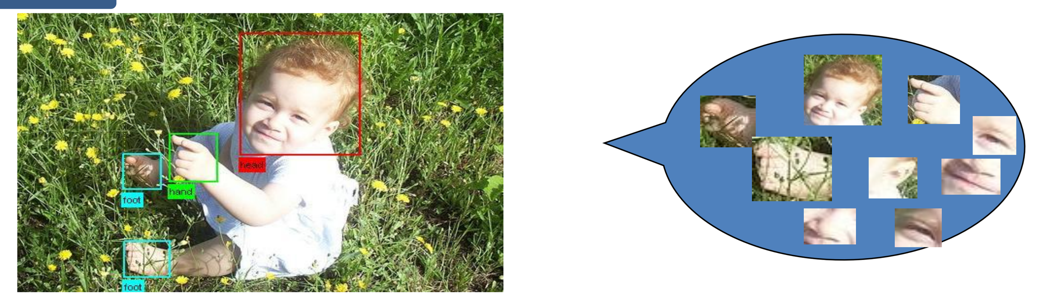
Sampling Strategies for Real-time Action Recognition

Feng Shi, Emil M. Petriu and Robert Laganière
School of EECS, University of Ottawa, Ontario, Canada

Introduction

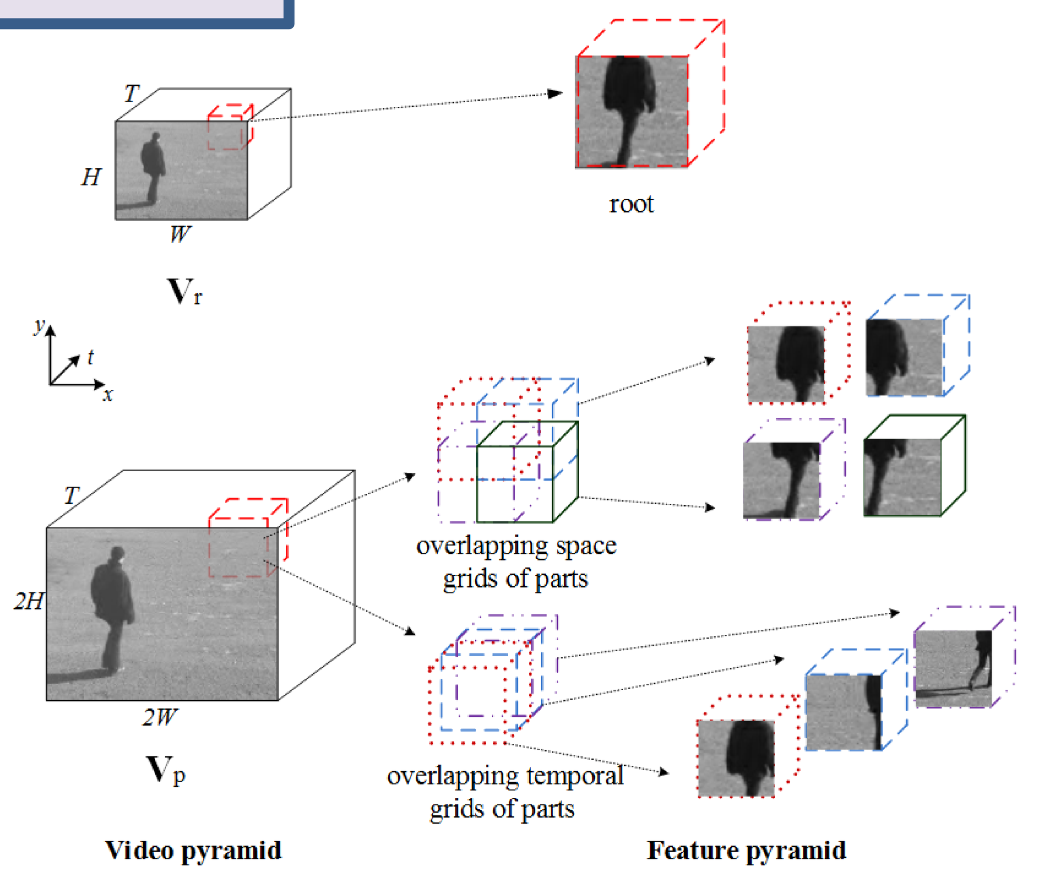


- Weakness of BoF approach:
  - Only containing statistics of unordered “features”: lost order – the arrangement of the set of events
  - Ignoring global information: lost spatial relationship
- Local spatio-temporal features are too sparse and expensive to extract
- The foreground actions in real-life video contain highly correlated background features

Diving with water background

Skiing with snow background

Local part model



root model

part model

- Dealing with out-of-order problem of BoF
  - Coarse “root” model containing local global information
  - High-resolution “part” models incorporating the temporal order information by including local overlapping “events”

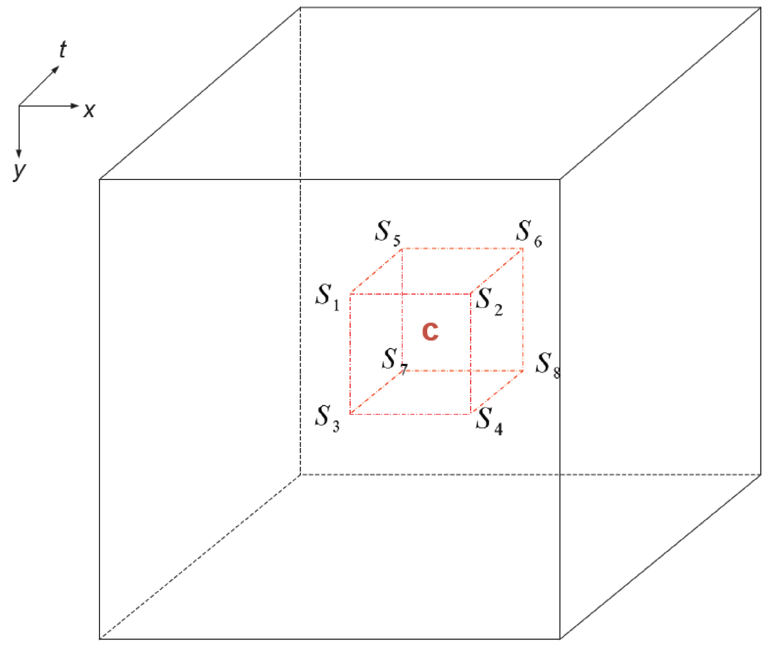
Sampling strategies

Spatial resolution	Cuboid [Dollár]	Dense [Wang]	Our sampling grid Samples/frame
80x60			767
160x120			4,295
360x288	44	643	27,950

- The very high dense sampling grid
  - Cubic patch: 1 root + multiple parts (1+8)
  - 72K features on a 80x60x94 video
  - Feature pyramid: 8 spatial scales, 2 temporal scales
- Increasing sampling density:
  - Sampling grid determined by “root” video at half the resolution
  - Decreasing sampling step size
  - Small initial patch size at 16x16x10
- Random sampling over sampling grid : 10K features (14%)
  - 10K roots + 80K parts

Efficiency

- Using integral video for fast cubic feature computation
  - Volume ***C*** can be computed with eight array references
- Two integral videos
  - 1 root + 1 part
  - Memory: 1(part)+0.25(root at half resolution) = 1.25 times video size
  - Efficiently computing root integral video by down-sampling part integral video
- No time spent on feature detection for random sampling
- Using FLANN (*Fast Approximate Nearest Neighbor Search Library*) for fast bag-of-words matching



$$S_8 - S_7 - S_6 + S_5 - S_4 + S_3 + S_2 - S_1$$

Descriptor	Feature size	Speed (frames per second)						Mean accuracy
		Integral video	Sampling	Flann BoF matching		Total fps		4k words
				4k words	6k words	4k words	6k words	
MBH	1152	41.19	192.4	267.14	252.78	30.79	29.92	41.1% ± 0.23
HOG3D	864	71.88	159.60	290.81	282.60	42.22	41.69	33.3% ± 0.19

Average computation speed with single core (i7-3770K)

Multi-channel SVM

$$K_{IH}(x_i, x_j) = \sum_c \frac{\omega^c}{\max(\omega^c)} \min(x_i^c, x_j^c)$$

- Novel method to combine multiple channels of different descriptors
  - Efficient histogram intersection kernel
  - More weight on discriminative descriptors

Comparison to state-of-the-art

Method		HMDB51	UCF50
HMDB51		23.2%	47.9%
ActionBank		26.9%	57.9%
MIP		29.17%	72.68%
Subvolume		31.53%	-
MRP		40.7%*	-
GIST3D		29.2%*	73.7%*
UCF50		27.02%*	76.90%*
Dense trajectories		46.6%*	**84.5%***
Ours	HOG	21.0%±0.28	58.6%±0.16
Ours	HOF	33.5%±0.31	69.7%±0.12
Ours	HOG3D	34.7%±0.40	72.4%±0.02
Ours	MBH	43.0%±0.11	80.1%±0.39
Ours	Combined	**47.6%**±0.29*	83.3%±0.15*

- Brute-force bag-of-words matching
- Intersection kernel LIBSVM one-verse-one
  - More weight on MBH descriptor
  - Better results with one-verse-all

Conclusion

- Very high density sampling without loss of efficiency
- More features, better performance
- Answered the out-of-order problem of bag-of-feature approach