Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition

Behrooz Mahasseni and Sinisa Todorovic

Oregon State University

CVPR 2016

Challenges of Large Scale Action Recognition

- Large number of action classes
  - Large variations within a single class
  - Small differences between distinct classes

Examples of **different** actions in the Sports-1M dataset

Downhill mountain biking

Road bicycle racing

Track cycling

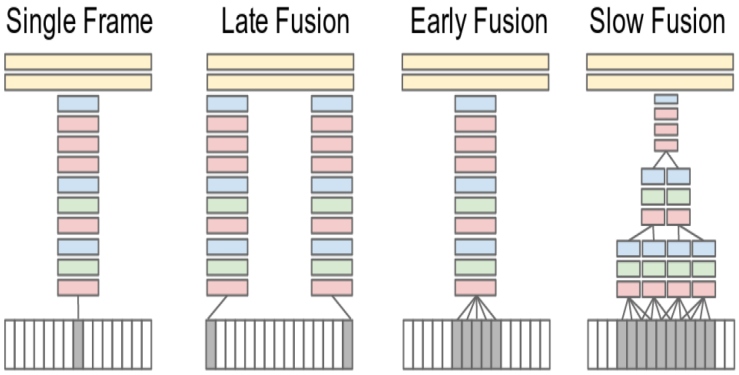
Challenges of Large Scale Action Recognition

- Wide range of camera views & motions

Examples of the volleyball action in the Sports-1M dataset

Recent Work

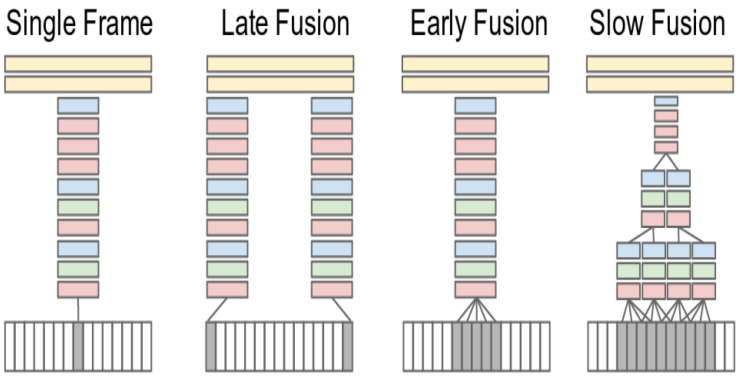
- Features are learned for classification
- Scalable, and transferable between domains
- Fast inference



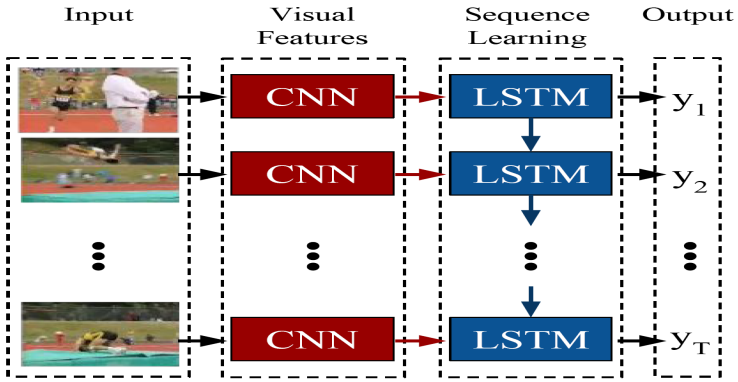
Karpathy et al., 2014
Ng et al., 2015
…

Recent Work

- Features are learned for classification
- Scalable, and transferable between domains
- Fast inference



Karpathy et al., 2014
Ng et al., 2015
…



Donahue,et al., 2014
Srivastava et al., 2015
…

Current Trends

Deeper Models

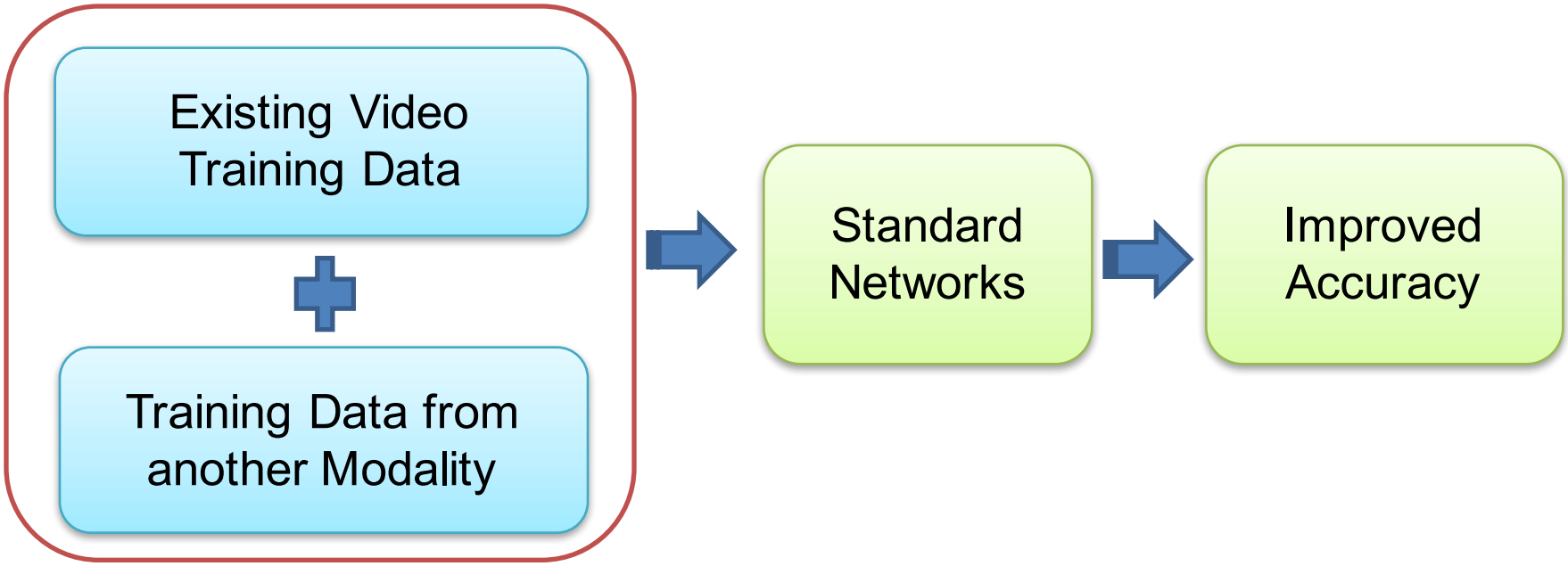
(e.g. Ng et al., 2015, Karpathy et al., 2014)

More Training Data

(e.g. Sports-1M, Activity Net)

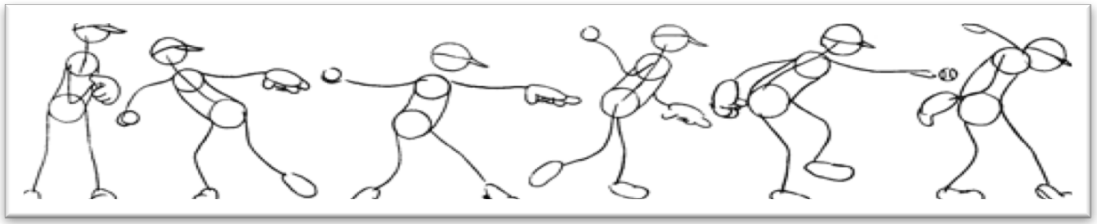
Our Key Idea

Use another modality with complementary information about human actions



Our Choice of Additional Modality

- Abstraction helps to understand complex concepts
- Sketches help to create abstract concepts



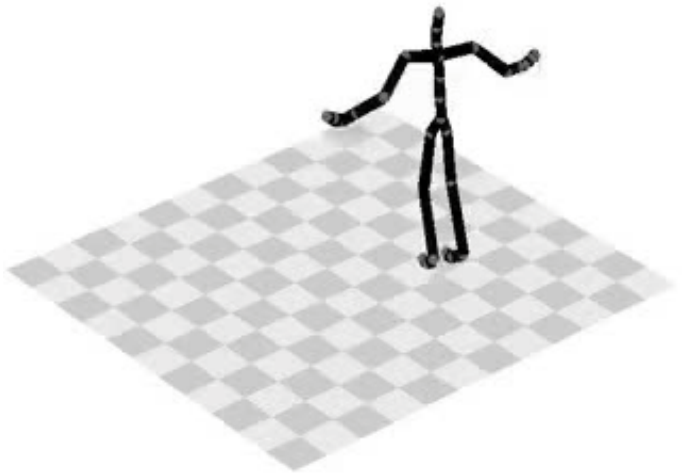
CogSci Lit:

- [1] Do Children Need Concrete Instantiations to Learn an Abstract Concept? [2006]
- [2] Abstraction processes during concept learning: A structural view [1988]
- [3] From Perceptual Categories to Concepts: What Develops? [2010]

3D Human Skeleton Sequences

- View-invariant and noise free
- Lower dimensional input space

Sports-1M videos,
Karpathy et al., 2014



HDM05

Limitations of 3D Skeleton Sequences

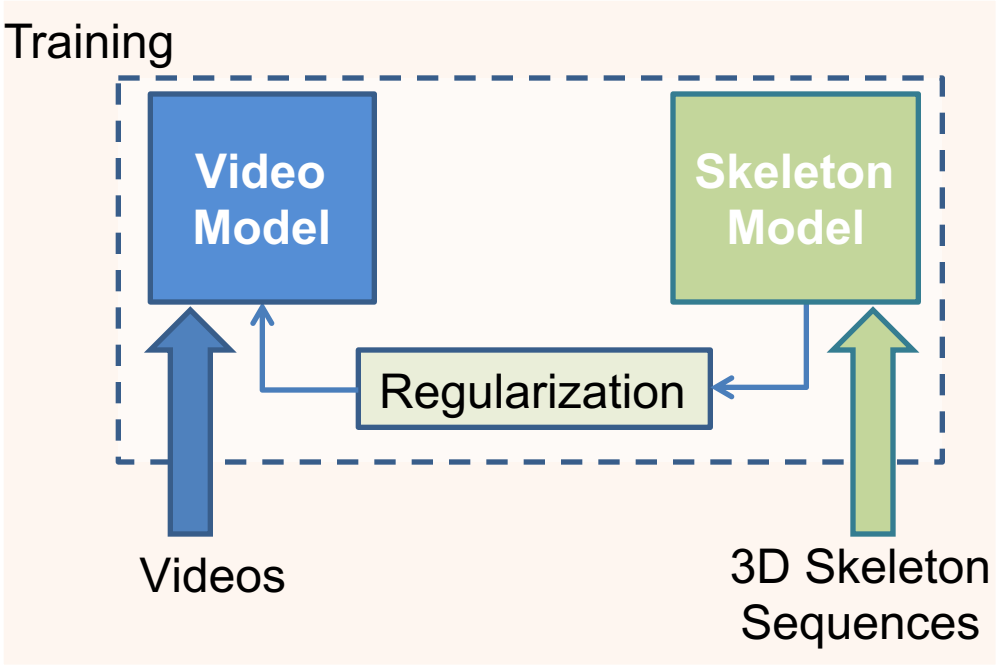
Poor Coverage of Action Classes

Most Skeleton Sequences represent Indoor Actions

Hard to Access at Test Time

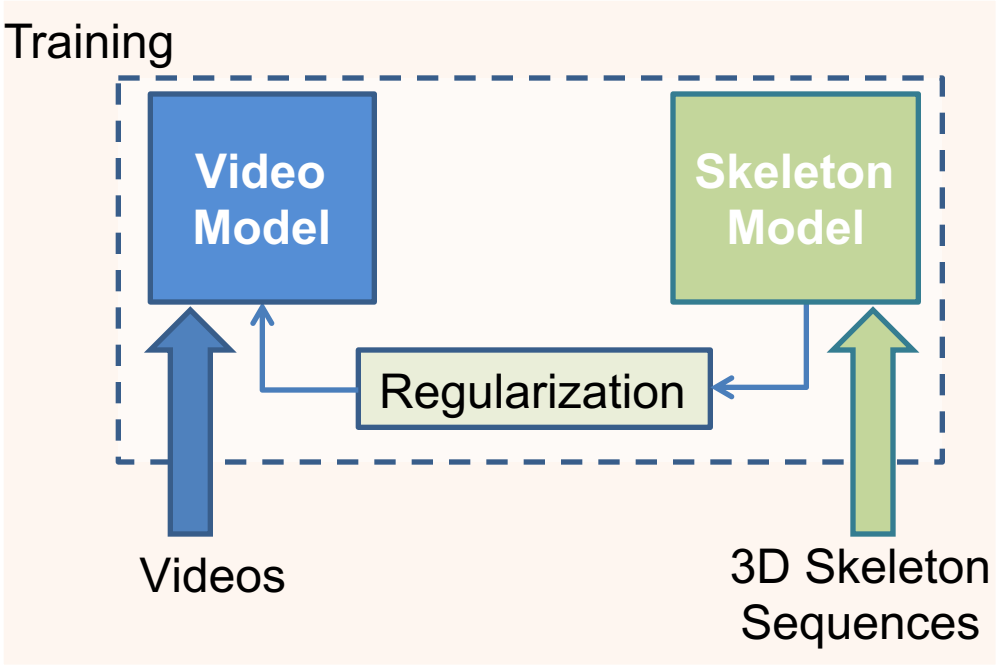
Our Approach

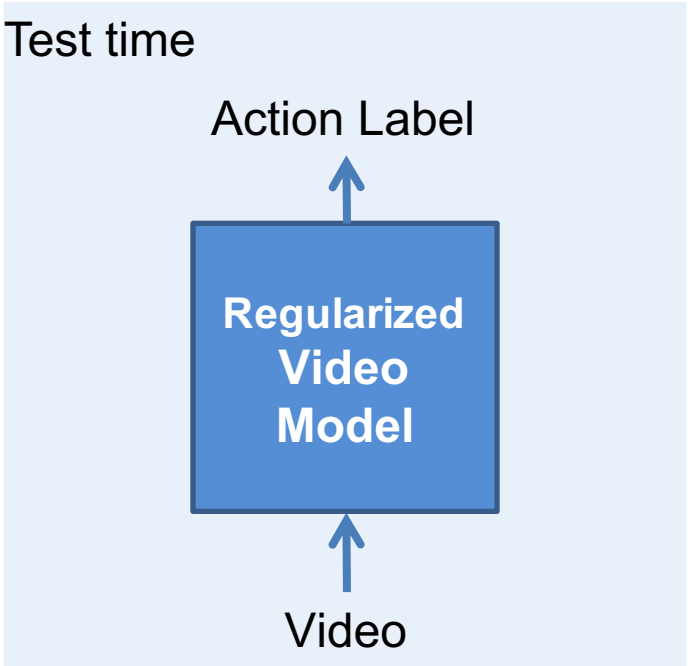
- Multimodal learning
- Regularized 2D video model



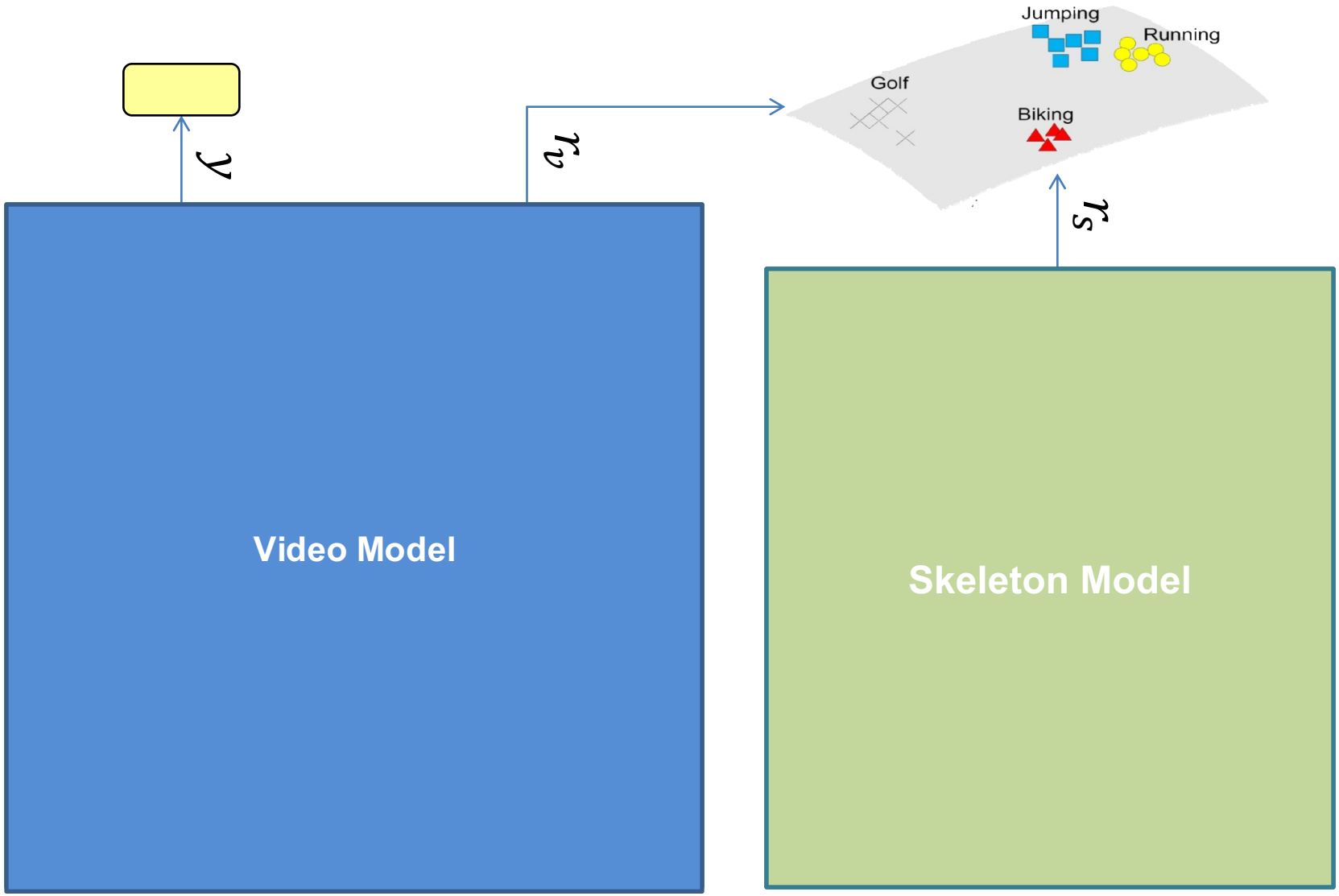
Our Approach

- Multimodal learning
- Regularized 2D video model

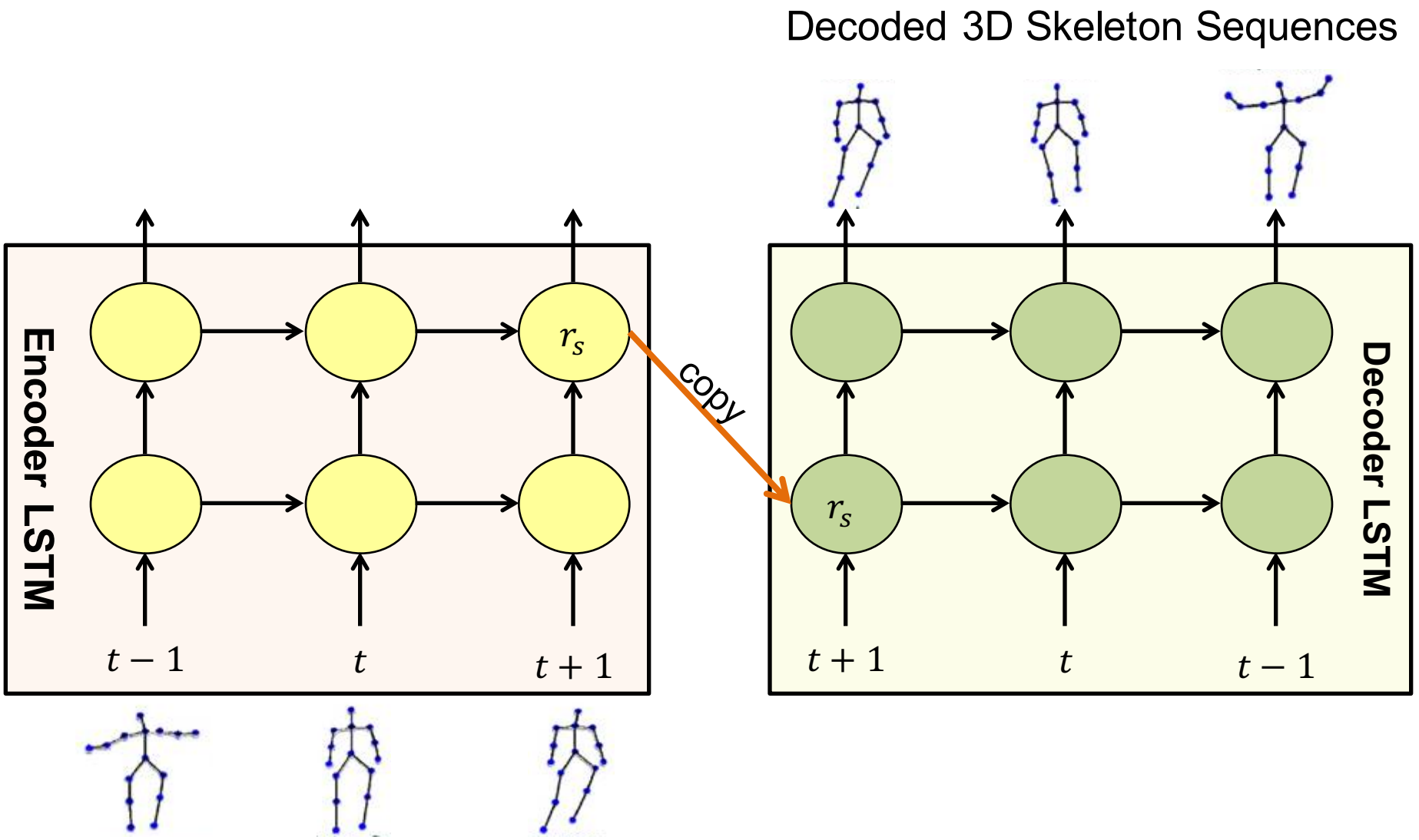




Training Framework

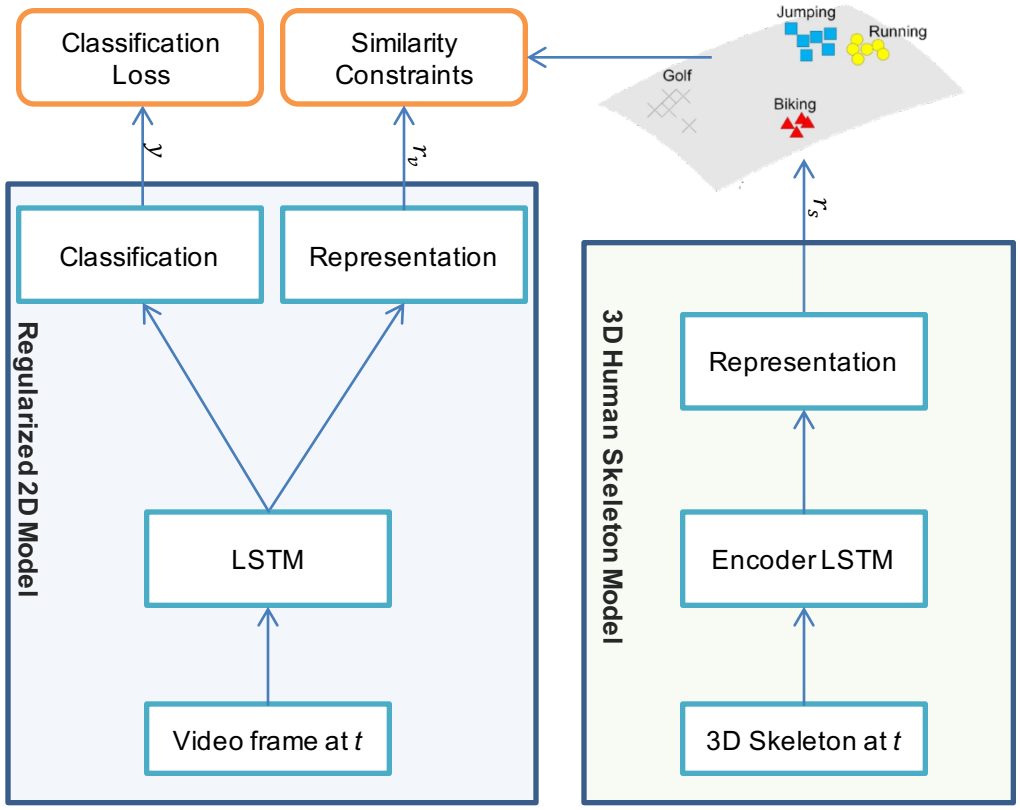


Encoder LSTM (e-LSTM)



Regularized Learning

- Classification loss
- Similarity constraints
  - Class independent
  - Class aware



Training

- Problem: Constraint optimization
- Solution: Hybrid backpropagation through time

$$\begin{cases} \text{If no condition is violated : } \Theta_t \leftarrow \Theta_{t-1} + \nabla Loss(\Theta) \\ \text{If any condition is violated: } \Theta_t \leftarrow \Theta_{t-1} + \sum \nabla Constraint(\Theta) \end{cases}$$

Constraints

- Class independent

$$\frac{1}{n}\sum_{r_s}|r_s - r_v| \leq \alpha$$

- Class aware

$$\underbrace{\frac{1}{n_1}\sum_{r_s}|r_s - r_v|}_{\text{Sum over same label instances}} - \overbrace{\frac{1}{n_2}\sum_{r_{s'}}|r_{s'} - r_v|}^{\text{Sum over different label instances}} \leq 0$$

Results

- Dataset: Sport1M

Method	Hit@1	Hit@5
Single Frame	59.3	77.7
LSTM	71.3	89.9
[1]	60.9	80.2
[2]	72.1	90.6
[3]	61.1	85.2
R-LSTM	**75.9**	**91.7**

[1] Karpathy et al. Large-scale video classification with convolutional neural networks. In CVPR, 2014
[2] Ng et al. Beyond short snippets: Deep networks for video classification, arXiv2015
[3] Tran et al. C3D: generic features for video analysis. CoRR 2014

Results

- Datasets:UCF101, HMDB-51

Method	UCF101	HMDB-51
[1]	65.4	-
[2]	75.8	44.1
[3]	71.12	-
[4]	72.8	40.5
[5]	79.34	-
[6]	85.2	-
R-LSTM	**86.9**	**55.3**

[1] Karpathy et al. Large-scale video classification with convolutional neural networks. CVPR, 2014
[2] Srivastava et al. Unsupervised learning of video representations using lstms, arXiv2015
[3] Donahue et al. Long-term recurrent convolutional networks for visual recognition and description, arXiv 2014
[4] Simonyan et al. Two-stream convolutional networks for action recognition in videos NIPS 2014
[5] Zha et al. Exploiting image-trained cnn architectures for unconstrained video classification, arXiv 2015
[6] Tran et al. C3D: generic features for video analysis, CoRR, 2014

Insights

- Actions that are directly about human motion

Actions	Accuracy Improvement
Running	4.2%
Badminton	1.8%
Track cycling	2.3%
Road bicycle racing	1.4%
Down hill biking	0.9%
bmx	0.8%

- Actions that are not about human motion

Actions	Accuracy Drop
Wind Surfing	-1.2%
Fishing	-1.0%
Land Surfing	-0.9%

Merit of adding 3D skeletons

- Accuracy vs Amount of training data

Training setup	Hit@1	Hit@5
100% 2D training data	71.3	89.9
99.5% 2D training data	71.2	89.9
99.5% 2D training data + 3D sequence data	**75.9**	**91.7**

Summary

- 3D sequences → Dynamics of human actions
- e-LSTM → Feature space from 3D sequences
- R-LSTM → Regularized video model
- Hybrid backpropagation through time
- Improved accuracy on benchmark datasets

Network Details

- DCNN + LSTM
  - Modified GoogLeNet
  - 2 Layer LSTM with 2048 and 1024 hidden units
  - Representation Layer output with 512 units
- e-LSTM
  - Input units = 54 (18*3)
  - 2 Layer LSTM with 1024and 512 hidden units