MULTI-MODAL SPATIO-TEMPORAL PYRAMID MATCHING FOR 3D HUMAN ACTION RECOGNITION

Charles Sturt University

Bin Liang (bliang@csu.edu.au)

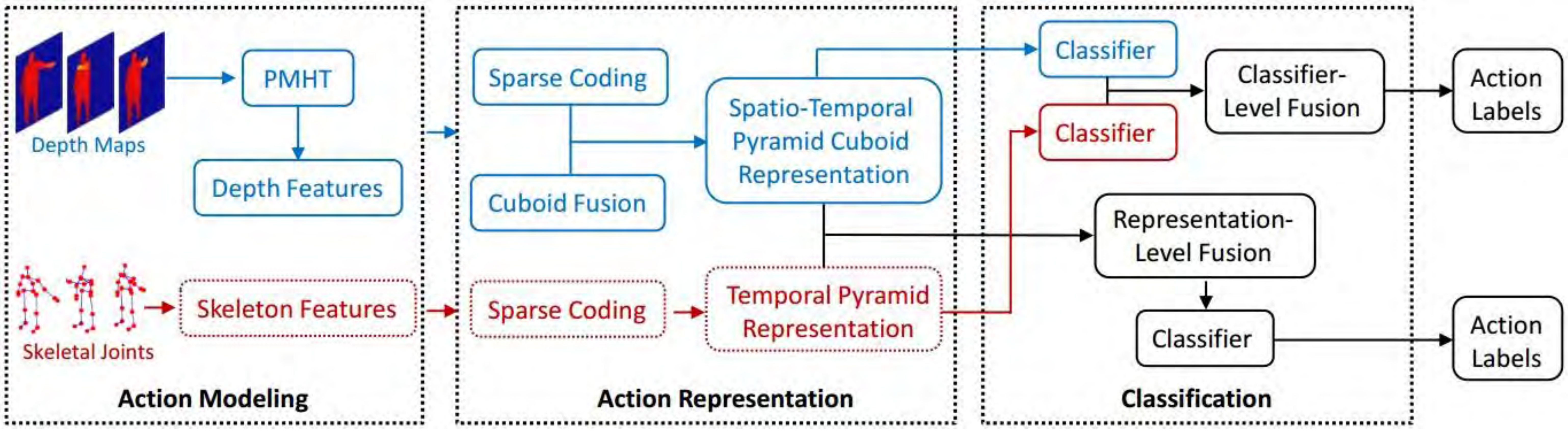
School of Computing and Mathematics, Charles Sturt University, Australia

Registration
Australian Workshop on Video/Image Coding, Processing, and Understanding
VICPU-2015

Motivations

- Complex human actions may have several temporal stages. In each stage, the temporal structural information is crucial to model complex actions.
- 3D locations of action motion encode strong discriminative information. Representing the 3D spatial information is quite challenging.
- A single modality is usually insufficient to solve the problem of recognizing human actions.

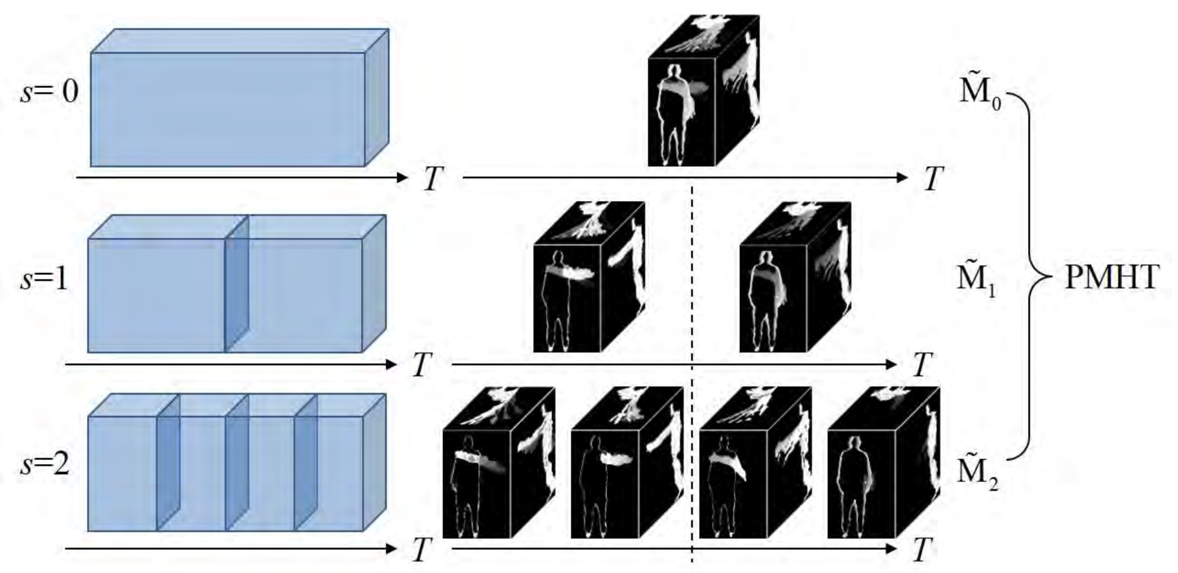
Framework



Depth-based Spatio-Temporal Pyramid Cuboid

Pyramid Motion History Templates (PMHT)

$$\mathrm{M}_v(P_v,t_1,t_2)=\begin{cases} t_2-t_1+1 & \text{if } D_v(P_v,t)>\xi \\ \max(0,\mathrm{M}_v(P_v,t-1)-\sigma) & \text{otherwise}\end{cases}$$



- Each depth frame is projected onto three orthogonal Cartesian planes.
- The projected sequence is partitioned into multi-scale sub-volumes.
- A depth sequence is represented by PMHT in multiple temporal scales.

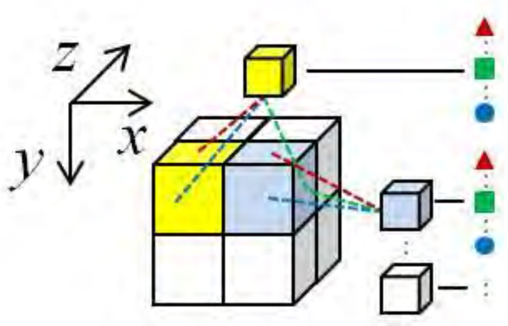
Depth Feature Extraction and Sparse Coding

- Only appearance-based features (HOG and SIFT) are used.
- Sparse Coding:

$$\{\mathbf{A}_v,\mathbf{D}_v\}=\underset{\mathbf{A}_v,\mathbf{D}_v}{\operatorname{argmin}}\left\{\sum_{i=1}^{n_v}\|\mathrm{x}_i^v-\mathbf{D}_v\alpha_i^v\|_2^2+\lambda\|\alpha_i^v\|_1\right\}$$

$$\text{s.t. } \|\mathrm{d}_j^v\|_2\leqslant 1,\quad \text{for } \forall j=1,\ldots,p_v$$

Spatio-Temporal Pyramid Cuboid Representation



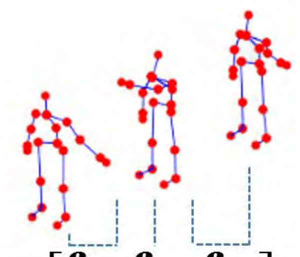
$$\tilde{F}^{(l)}=[\underbrace{\mathbf{z}_{\hat{f}}^{(l)}(1,1),\mathbf{z}_{\hat{t}}^{(l)},\mathbf{z}_{\hat{s}}^{(l)}}_{cuboid\ 1},\ldots,\underbrace{\mathbf{z}_{\hat{f}}^{(l)}(2^l,2^l),\mathbf{z}_{\hat{t}}^{(l)},\mathbf{z}_{\hat{s}}^{(l)}}_{cuboid\ 2^l\times 2^l\times 2^l}]$$

$$=\mathop{\Big\|}_{i,j,k=1}^{2^l}[\mathbf{z}_{\hat{f}}^{(l)}(i,j),\mathbf{z}_{\hat{t}}^{(l)}(k,j),\mathbf{z}_{\hat{s}}^{(l)}(i,k)]$$

- ***Cuboid fusion*** combines the spatial-dependent sparse codes from the corresponding grids on three planes to construct the cuboid representation ($\|$ denotes vector concatenation).

Skeleton-based Temporal Pyramid

Skeleton Feature Extraction and Sparse Coding



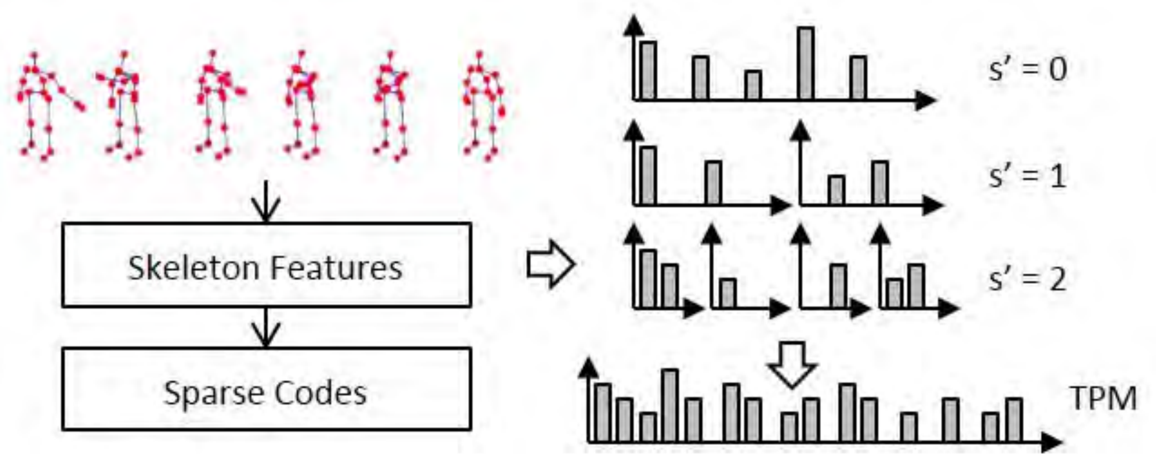
$$\mathrm{f}_c=[\mathbf{f}_{ci},\mathbf{f}_{cc},\mathbf{f}_{cp}]$$

$$\mathbf{f}_{ci}=\{\bar{\mathrm{p}}_i(c)-\bar{\mathrm{p}}_j(i)\,|\,\bar{\mathrm{p}}_i(c)\in\bar{\mathbf{p}}(c);\bar{\mathrm{p}}_j(i)\in\bar{\mathbf{p}}(i)\}$$

$$\mathbf{f}_{cc}=\{\bar{\mathrm{p}}_i(c)-\bar{\mathrm{p}}_j(c)\,|\,i,j=1,\ldots,N_s;i\neq j\}$$

$$\mathbf{f}_{cp}=\{\bar{\mathrm{p}}_i(c)-\bar{\mathrm{p}}_j(p)\,|\,\bar{\mathrm{p}}_i(c)\in\bar{\mathbf{p}}(c);\bar{\mathrm{p}}_j(p)\in\bar{\mathbf{p}}(p)\}$$

Temporal Pyramid Representation (TPM)



- The temporal information is well kept by temporal segments.
- TPM is not sensitive to the temporal shift or misalignment since lower level of the pyramid keeps less temporal information.

Multi-Modal Fusion

- **Representation-level Fusion**: the representations generated from two modalities are concatenated together to form a final representation as the input to the classifier.
- **Classifier-level Fusion**: the classifiers for two modalities are trained separately and classifier combination is performed subsequently to generate the final result.
  - Arithmetic Mean (AM), Geometric Mean (GM)
  - Logistic Regression (LR): learn weights for the combination.

Experimental Results

MSR Action 3D Dataset

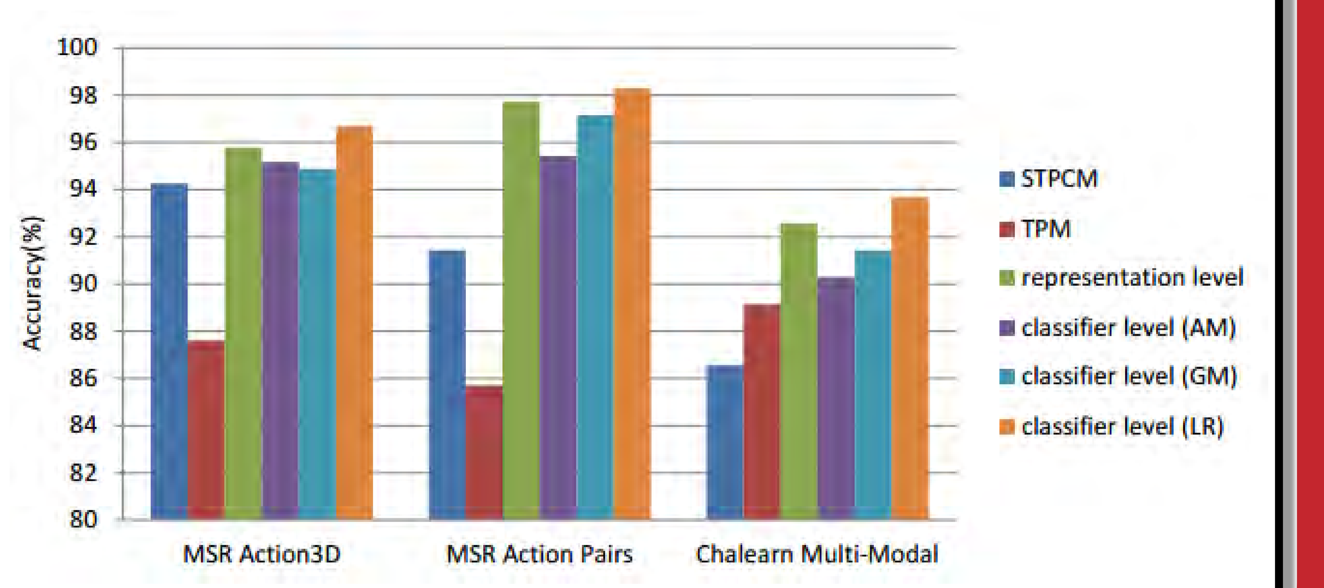
Method	Modality	Accuracy
Bag of 3D points [23]	depth	74.70
HOJ3D [46]	skeleton	78.97
Eigenjoints [48]	skeleton	82.33
STOP [39]	depth	84.80
ROP [40]	depth	86.20
Actionlet Ensemble [42]	depth + skeleton	88.20
HON4D [34]	depth	88.89
DSTIP [44]	depth	89.30
3DMTM-PHOG [25]	depth	90.70
DMM-HOG [50]	depth	91.70
SNV + joint trajectories [49]	depth + skeleton	93.09
STPCM (normal fusion)	depth	**92.45**
STPCM (cuboid fusion)	depth	**94.26**
TPM	skeleton	**87.61**
STPCM + TPM	depth + skeleton	**96.68**

2014 Chalearn Multi-Modal Dataset

Method	Modality	Accuracy
2DMTM [24]	depth	76.99
Multi-modality Recognition [10]	RGB+depth+skeleton	90.30
Skeleton + 2DMTM [24]	skeleton + depth	92.80
STPCM (normal fusion)	depth	**81.85**
STPCM (cuboid fusion)	depth	**86.56**
TPM	skeleton	**89.14**
STPCM + TPM	depth + skeleton	**93.67**

MSR Action Pairs Dataset

Method	Modality	Accuracy
Actionlet Ensemble [42]	depth + skeleton	63.33
DMM-HOG [50]	depth	66.11
Actionlet Ensemble + Pyramid [42]	depth + skeleton	82.22
HON4D [34]	depth	96.67
SNV [49]	depth + skeleton	98.89
STPCM (normal fusion)	depth	**88.57**
STPCM (cuboid fusion)	depth	**91.43**
TPM	skeleton	**85.71**
STPCM + TPM	depth + skeleton	**98.29**



Evaluation of various multi-modal fusion schemes