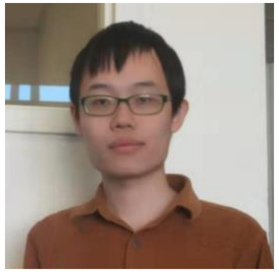
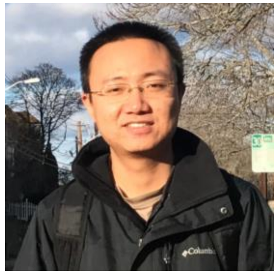
Generative View-Correlation Adaptation for Semi-Supervised Multi-View Learning



Yunyu Liu

liu.yuny@northeastern.edu



Lichen Wang

wanglichenxj@gmail.com

Yue Bai

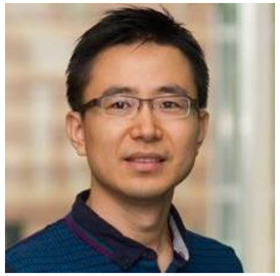
bai.yue@northeastern.edu

Can Qin

qin.ca@norteasterh.edu

Zhengming Ding

zd2@iu.edu



Yun Fu

yunfu@ece.neu.edu

SMILE Lab

Electrical & Computer Engineering

Northeastern University

SmileLab website: https://web.northeastern.edu/smilelab/
Code will be available: https://github.com/wenwen0319/GVCA

Introduction

Topic:

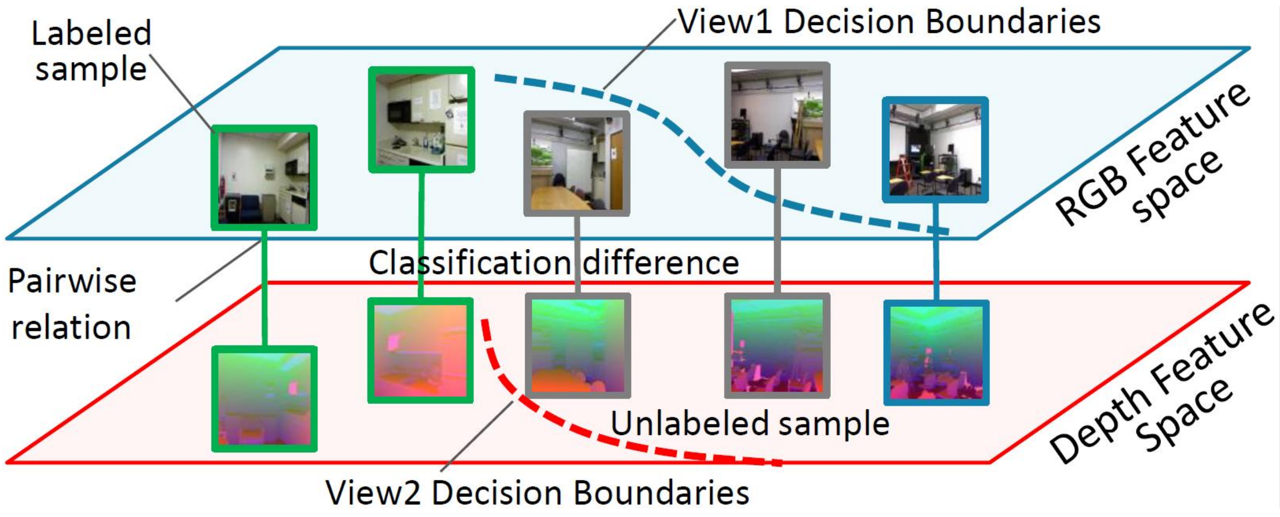
- Multi-view Action Recognition

Setting:

- Input: Labeled and unlabeled Multi-view action sequences (e.g., RGB + Depth)
- Output: Action prediction

Challenges:

- Heterogeneous multi-view feature domains
- Small dataset; hard to label
- Inconsistent view-specific predictions

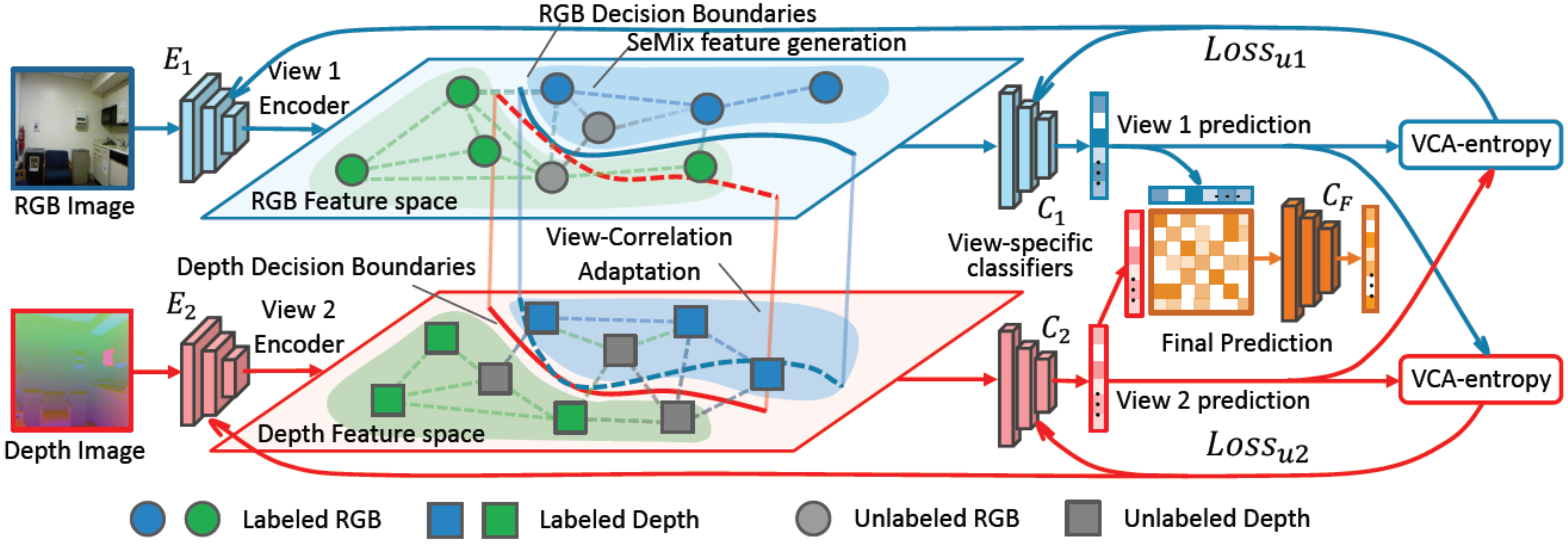


Concept of Multi-view Action Recognition

Our approach

Generative View-Correlation Adaptation for Semi-Supervised Multi-View Learning (GVCA)

1. A novel fusion strategy named View-Correlation Adaptation (VCA) is deployed in both feature and label space.
2. A new SeMix approach to generate samples using both labeled and unlabeled data.
3. An effective label-level fusion network is proposed to obtain the final classification result.

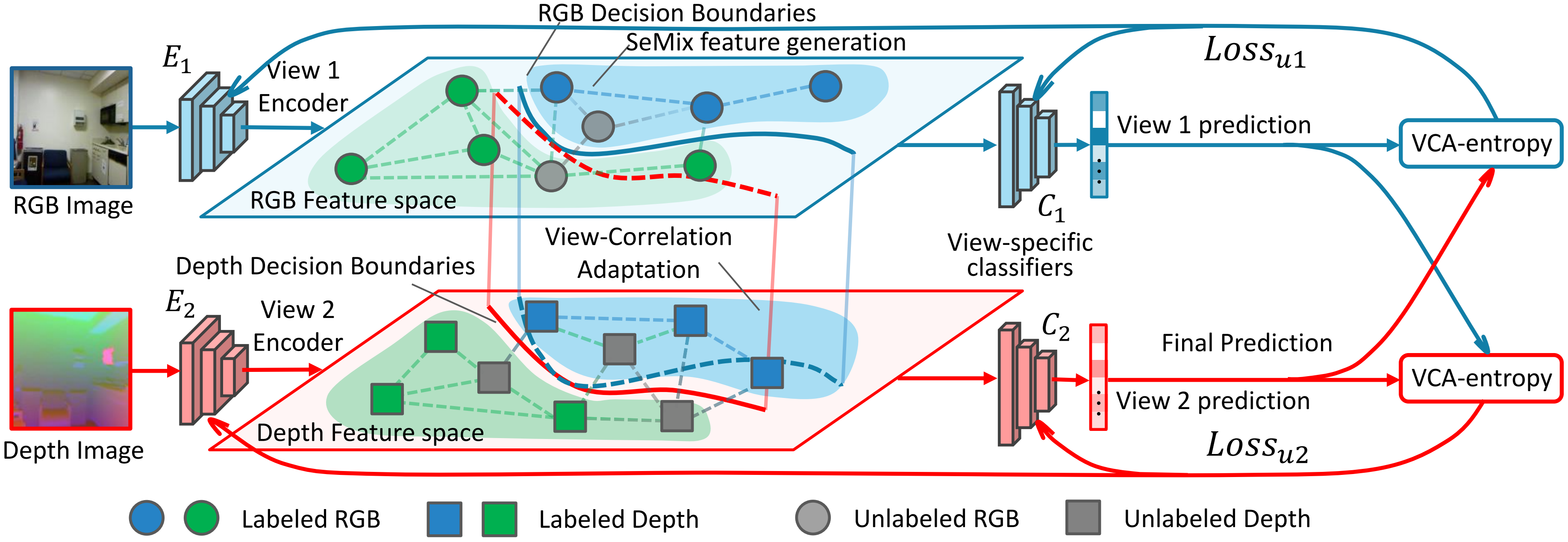


Framework of GVCA

Our approach

View-Correlation Adaptation and Entropy-based version

1. A novel fusion strategy named View-Correlation Adaptation (VCA) is deployed in both feature and label space.
2. A new SeMix approach to generate samples using both labeled and unlabeled data.
3. An effective label-level fusion network is proposed to obtain the final classification result.



Framework of GVCA

SeMix

Mixup:

- The dataset has little labeled data. To fully explore the data, a data generating method is used.
- x_i, x_j are labeled features y_i, y_j are corresponding labels
- The new data can be generated by
  - $X = \alpha x_i + (1-\lambda)\, x_j \quad Y = \alpha y_i + (1-\lambda)\, y_j$

SeMix:

Insight: We explore the connections from both labeled and unlabeled samples.

$$X_{U\_i} = \lambda x_U + (1-\lambda)\, x_i$$
$$X_{U\_j} = \lambda x_U + (1-\lambda)\, x_j$$
$$Y_{U\_i} = \lambda y_U + (1-\lambda)\, y_i$$
$$Y_{U\_j} = \lambda y_U + (1-\lambda)\, y_j$$
$$\lambda' \sim Beta(\alpha, \alpha) \quad \alpha = 0.5$$
$$\lambda = max(\lambda', 1-\lambda')$$

Associate labeled + unlabeled samples

$$Loss = \left| \left(C(X_{U\_i}) - C(X_{U\_j}) \right) - (1-\lambda)(y_i - y_j) \right|$$

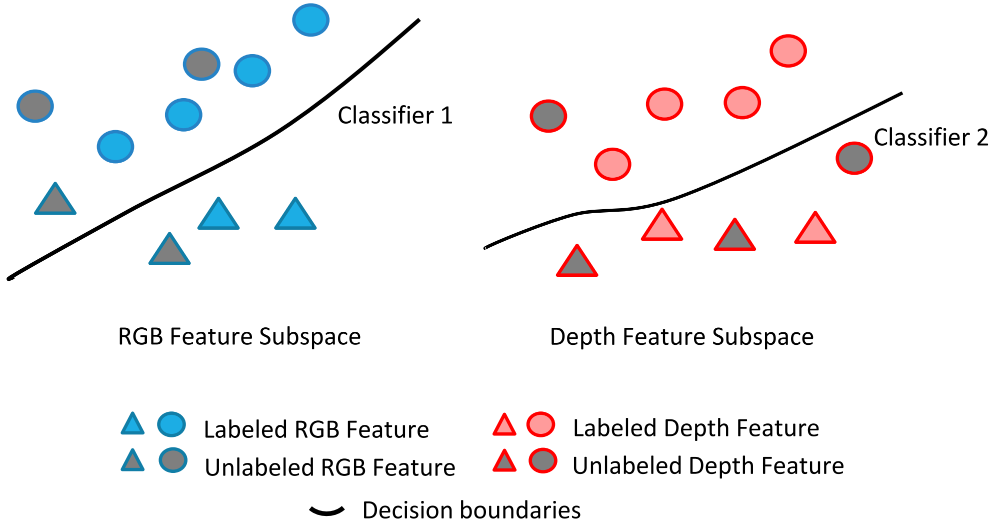
Label consistency loss

where C is the classifier, x_U is the feature of an unlabeled data.

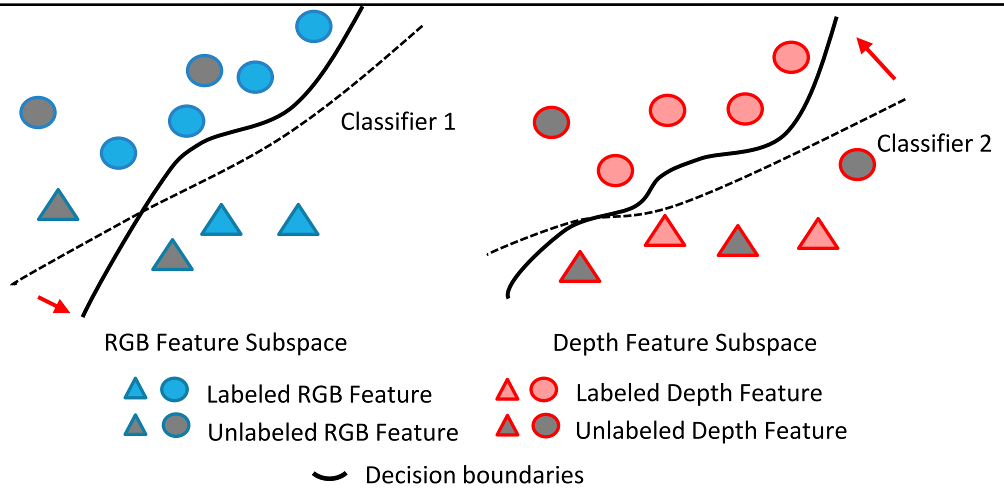
VCA

Motivation:

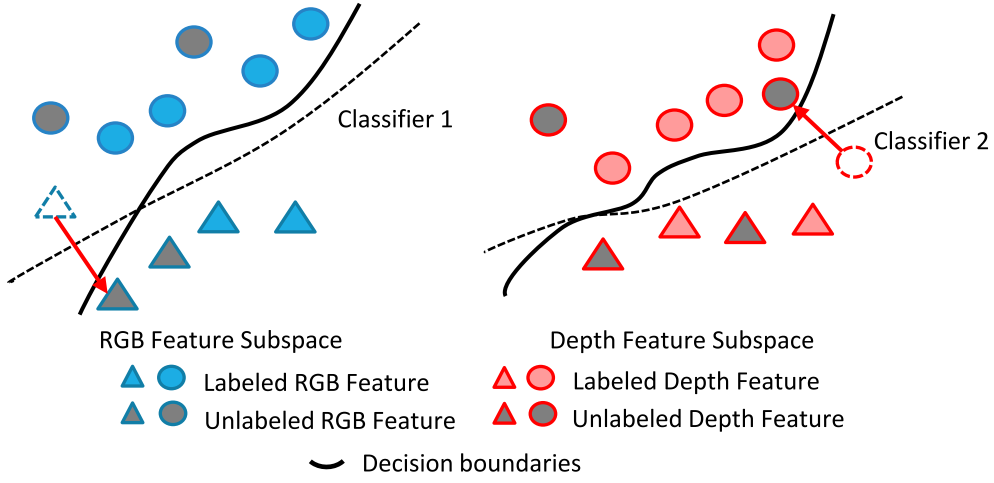
- **Inter-view adaptation**: Adapt the point representation based on classification guidance of another view



Step 1



Step 2



Step 3

VCA

Method:

- **Step 1 Update E_1 , E_2 , C_1 and C_2** with the $L_{labeled}$
- **Step 2**

  $\max\limits_{C_1,C_2} \boldsymbol{L_{unlabeled}}$, **fixed E_1 , E_2** .
- **Step 3**

  $\min\limits_{E_1,E_2} \boldsymbol{L_{unlabeled}}$, **fixed $C_1(.)$ and $C_2(.)$**.

- The pairwise loss of labeled features is

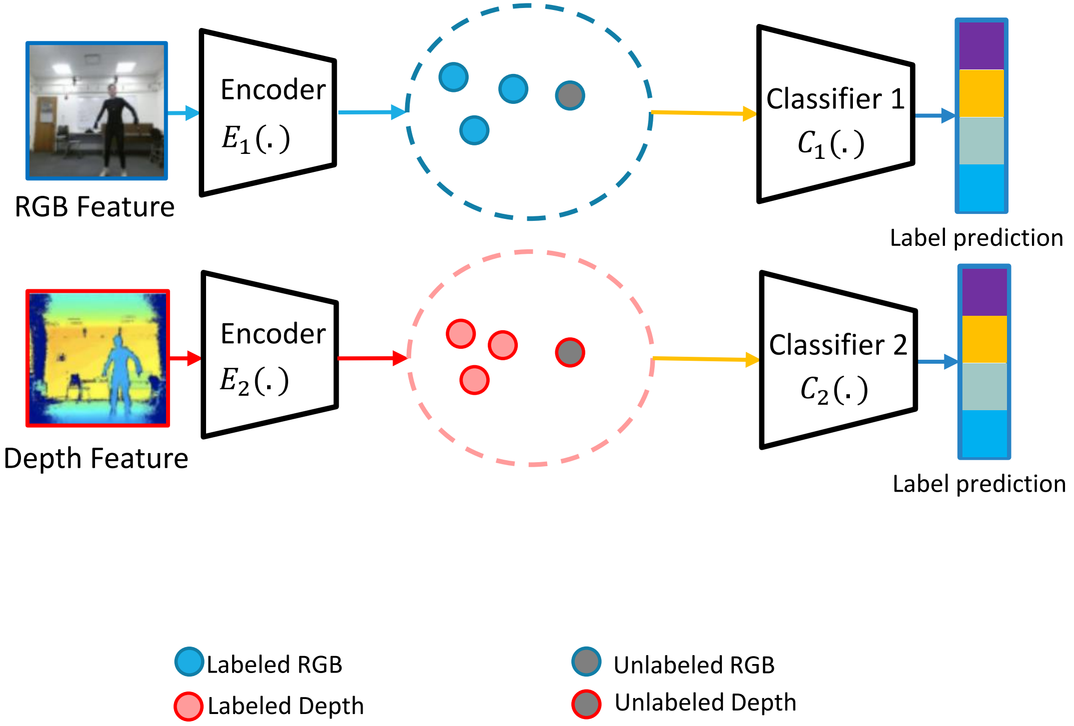
$$L_{labeled} = \sum_{i=1}^{2} \left\| y_j, C_i\left(E_i(x_j^i)\right) \right\|$$

  Where $\|.\|$ is L2 Normalization

- The pairwise loss of unlabeled features is

$$L_{unlabeled} = W\left(C_1(E_1(x_U^1)),\ C_2\left(E_2(x_U^2)\right)\right)$$

  Where W is Wasserstein Distance



VCA-entropy

Method:

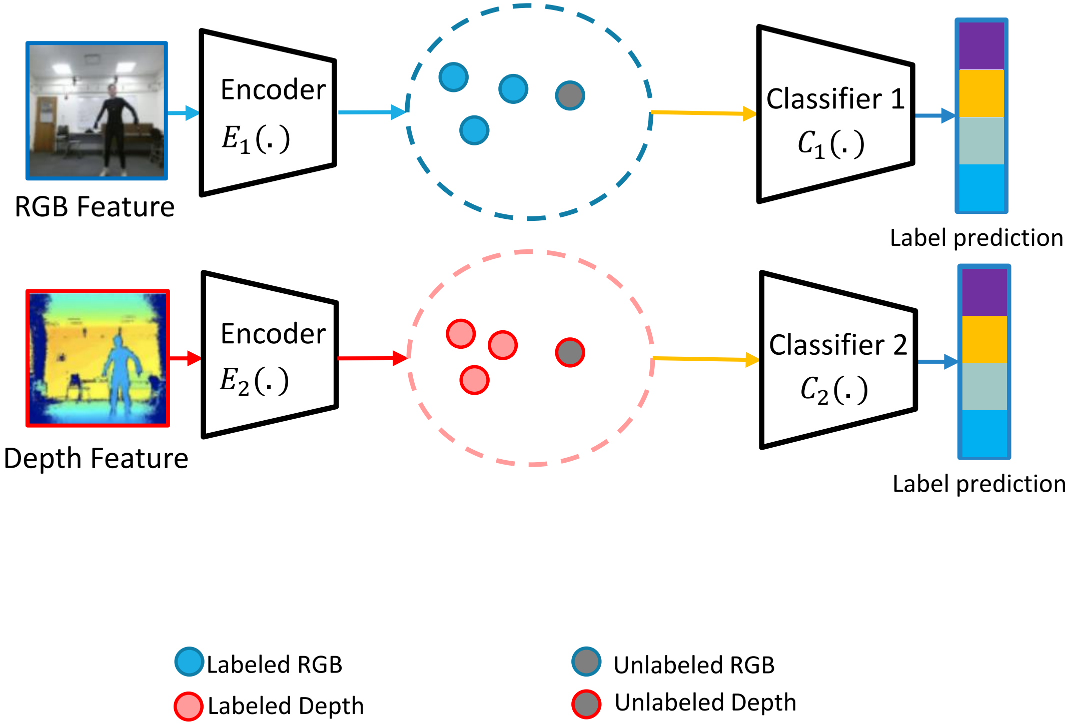
- **Step 1 Update E_1 , E_2 , C_1 and C_2** with the label loss of labeled features
- **Step 2**

  $\max_{C_1} L^1_{unlabeled}$, **fixed E_1.**

  $\max_{C_2} L^2_{unlabeled}$, **fixed E_2 .**
- **Step 3**

  $\min_{E_1} L^1_{unlabeled}$, **fixed C_1.**

  $\min_{E_2} L^2_{unlabeled}$, **fixed C_2.**



- H(.) is the entropy of a distribution $H(.) = -\sum p(x) \ln p(x)$
- $C_1(.) = C_1\left(E_1(x_U^1)\right)$ $C_2(.) = C_2\left(E_2(x_U^2)\right)$
- The pairwise loss of unlabeled features is

$$L^1_{unlabeled} = \frac{H\left(C_1(.)\right)}{H(C_2(.))} ||C_1(.) - C_2(.)||$$

$$L^2_{unlabeled} = \frac{H\left(C_2(.)\right)}{H(C_1(.))} ||C_1(.) - C_2(.)||$$

label-level fusion network

Insight:

The discrepancy still exists even after the alignment procedure.

We deploy a novel fusion strategy in label space.

$$\boldsymbol{L_F} = \left\| \boldsymbol{C_F}\left(\boldsymbol{reshape}(\tilde{\boldsymbol{y}}_1^T\tilde{\boldsymbol{y}}_1 + \tilde{\boldsymbol{y}}_1^T\tilde{\boldsymbol{y}}_2 + \tilde{\boldsymbol{y}}_2^T\tilde{\boldsymbol{y}}_2)\right) - \boldsymbol{y} \right\|_{\boldsymbol{F}},$$

Where $\tilde{y}_1$ and $\tilde{y}_2$ are initial results from view 1 and view 2, y is the ground truth of corresponding labeled data.

Experiments

Synergetic Media Learning Lab

Setting:

- Datasets: UWA[1], MHAD[2], and DHA[3]
- Multi-view action recognition
- Baselines and Performance. Classification accuracy (%)

Conclusion:

- High performance
- Effectiveness of all modules

Setting	Method	DHA	UWA	MHAD
RGB	LSR	65.02	67.59	96.46
	SVM [25]	66.11	69.77	96.09
	VLAD [11]	67.85	71.54	97.17
	TSN [34]	67.85	71.01	97.31
Depth	LSR	82.30	45.45	47.63
	SVM [25]	78.92	34.92	45.39
	WDMM [1]	81.05	46.58	66.41
RGB+D	LSR	77.36	68.77	97.17
	NN	86.01	73.70	96.88
	SVM [25]	83.47	72.72	96.80
	AMGL [20]	74.89	68.53	94.70
	MLAN [19]	76.13	66.64	96.46
	AMUSE [21]	78.12	70.32	97.23
	GMVAR [31]	88.72	76.28	98.94
	Ours	**89.31**	**77.08**	**98.94**

Classification Accuracy

Setting	RGB	Depth	RGB+D
TSN [34]	67.85	-	-
WDMM [1]	-	81.05	-
MLP	77.10	79.01	79.12
Mixup	68.51	81.43	81.48
SeMix	69.37	**82.73**	83.15
VCA	75.26	80.86	81.32
VCA-entropy	**80.86**	82.61	84.10
Ours complete	-	-	**89.31**

Ablation Study

[1] Hossein Rahmani, et al. Histogram of oriented principal components for cross-view action recognition.IEEE Trans. PAMI, 38(12):2430–2443, 2016
[2] Ferda Ofli, et al. Berkeley mhad: A comprehensive mul-timodal human action database. In Proc. IEEE WACV, pages53–60, 2013.
[3] Yan-Ching Lin, et al. Human action recog-nition and retrieval using sole depth information. In Proc.ACM MM, pages 1053–1056, 2012.

Experiments

Setting:

- Datasets: UWA[1], MHAD[2], and DHA[3]
- Multi-view action recognition
- Different ratios of labeled training samples and generate number

Conclusion:

- High performance when using less labeled data, achieves a comparable result using 50%.
- Achieves the peak at 1x and then fluctuates.

Table 3. Classification accuracy(%) given different ratios of labeled training samples.

Dataset	Ratio	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
DHA	GMVAR [31]	44.86	62.55	69.14	73.25	77.37	80.66	84.36	84.36	86.01	88.72
	Ours	48.15	69.14	74.90	79.42	83.54	83.56	85.60	86.83	87.24	89.31
UWA	GMVAR [31]	35.18	46.64	54.15	58.57	65.61	69.17	73.91	75.49	76.28	76.28
	Ours	36.36	57.71	62.85	64.03	67.98	70.75	73.12	76.56	76.66	77.08
UCB	GMVAR [31]	53.36	72.79	90.11	92.64	93.41	94.76	95.91	95.49	96.28	98.94
	Ours	52.30	75.83	91.17	92.23	92.93	93.11	95.05	96.82	97.88	98.94

Table 4. Accuracy (%) of different generate number

Dataset	0x	0.1x	0.3x	0.5x	1x	2x	3x
DHA	84.10	87.24	88.07	88.07	89.31	88.89	89.31
UWA	75.49	76.30	77.08	76.68	77.08	77.47	76.28
MHAD	98.23	98.59	98.23	98.23	98.94	98.94	97.88

[1] Hossein Rahmani, et al. Histogram of oriented principal components for cross-view action recognition.IEEE Trans. PAMI, 38(12):2430–2443, 2016
[2] Ferda Ofli, et al. Berkeley mhad: A comprehensive mul-timodal human action database. In Proc. IEEE WACV, pages53–60, 2013.
[3] Yan-Ching Lin, et al. Human action recog-nition and retrieval using sole depth information. In Proc.ACM MM, pages 1053–1056, 2012.

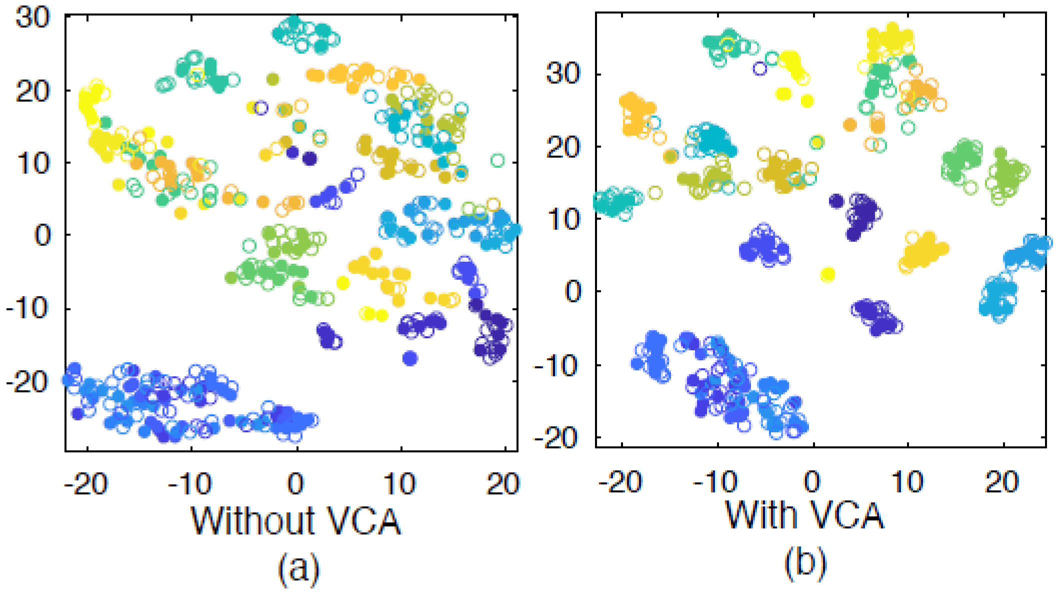
Experiments

Setting:

- Datasets: UWA[1], MHAD[2], and DHA[3]
- Multi-view action recognition
- TSNE visualization

Conclusion:

- High performance when using less labeled data, achieves a comparable result using 50%.



[1] Hossein Rahmani, et al. Histogram of oriented principal components for cross-view action recognition.IEEE Trans. PAMI, 38(12):2430–2443, 2016
[2] Ferda Ofli, et al. Berkeley mhad: A comprehensive mul-timodal human action database. In Proc. IEEE WACV, pages53–60, 2013.
[3] Yan-Ching Lin, et al. Human action recog-nition and retrieval using sole depth information. In Proc.ACM MM, pages 1053–1056, 2012.

Thank you!

Please contact: liu.yuny@northeastern.edu for questions.

SMILE Lab

Electrical & Computer Engineering

Northeastern University