EFFECT OF WAVELET AND HYBRID CLASSIFICATION ON ACTION RECOGNITION

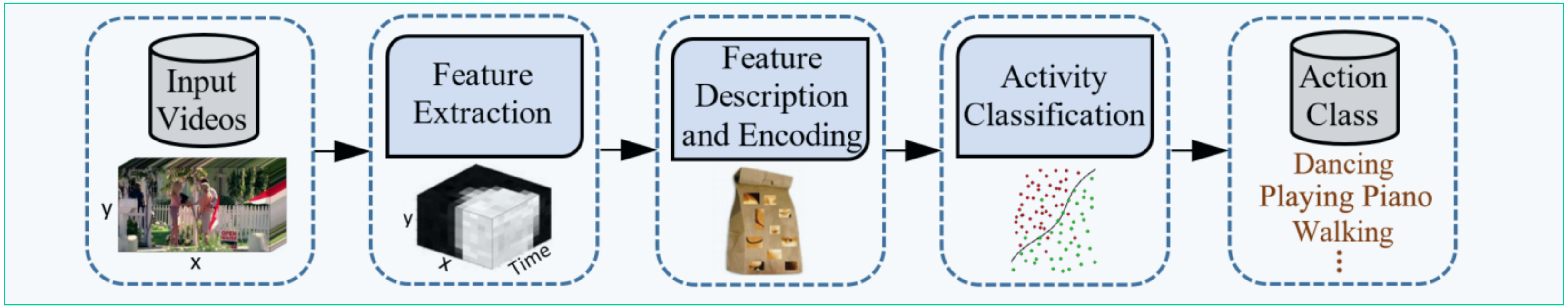
Eman Mohammadi
Q. M. Jonathan Wu
Yimin Yang
Mehrdad Saif

Computer Vision and Sensing Systems Laboratory

Department of Electrical and Computer Engineering, University of Windsor, Ontario, Canada

Introduction

- The bag of visual word framework leads to successful action recognition frameworks.



- Much less research has been performed on the preprocessing and classification stages.
- Action classification is tremendously challenging for computers due to the complexity of video data and the subtlety of human actions.

Introduction

- **Classification Step:** equivalent probabilities may be provided for running, jogging and walking classes while classifying the samples of **KTH dataset**.

Jogging

Running

Walking

- The classifier is not capable of making the final decision indubitably when equivalent probabilities are generated for different classes.

Contributions

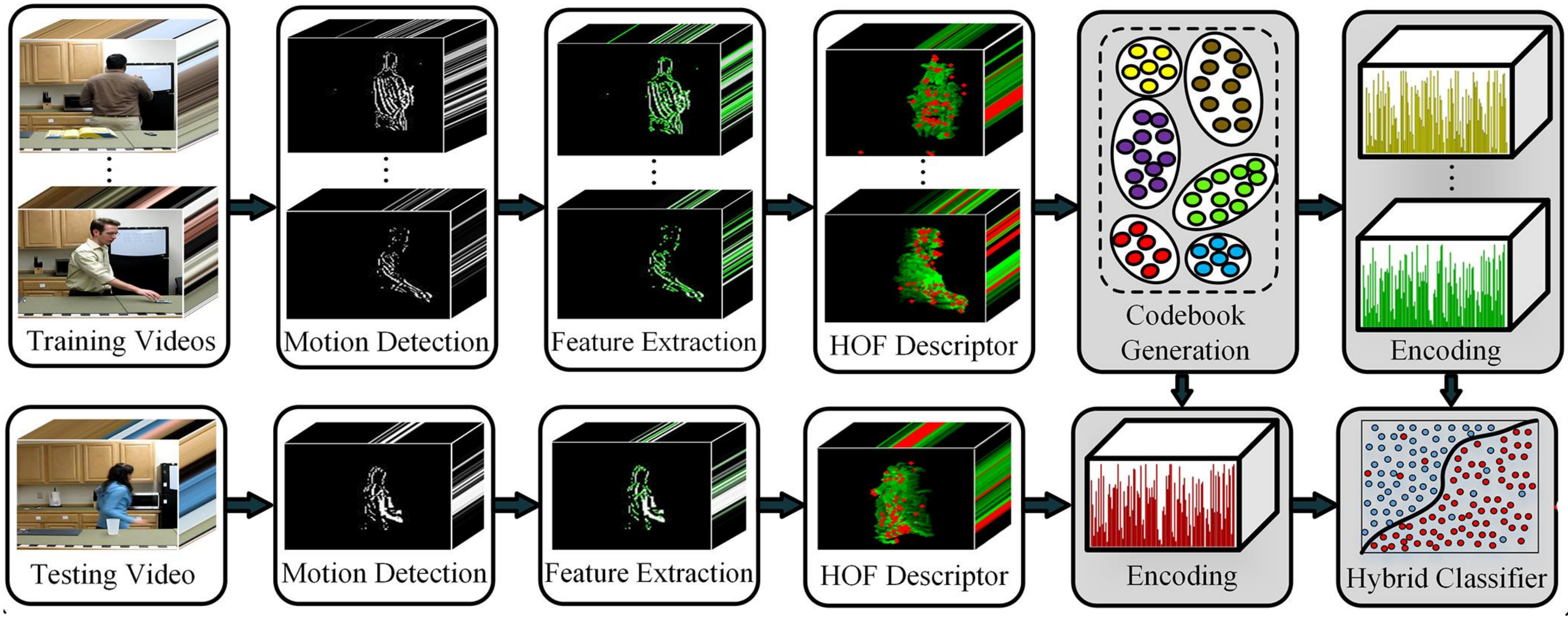
- **Classification Step:** Proposing a hybrid classifier **(including 3 layers)** to automatically compress the extracted features and select the best SVM kernel for action classification.

- Different dimensions are evaluated to optimize the compression rate in the 2nd layer of hybrid classifier.

- **Pre-processing Step:** we employ 3D-discrete wavelet transform (3D-DWT) to segment the moving objects in videos before local feature extraction.

- Different thresholding values are evaluated to extract the best motion saliency map for local feature extraction. The effect of 3D-DWT on motion-based features is evaluated in this paper.

Action Recognition Framework using Preprocessing and Hybrid Classification Steps



Motion Saliency Detection

- **3D Discrete Wavelet Transform (3D-DWT)** consists of three 1D-DWT in the x, y, and t directions.
- It is composed of **high-pass** and **low-pass filters** that perform a convolution of filter coefficients on input frames.
- **The output of 3D-DWT:** 8 sub-signals in three directions.
- We utilize the sub-signal which is generated by high-pass filter to each direction.

Steps to create motion saliency maps

1. Resize frames to 500x500 pixels
2. Apply 3D-DWT on the resized video frames
3. Create the transformed videos with 10 frames per second
4. Utilize the threshold of 200 to make the binary videos including motion saliency maps.

Feature Extraction

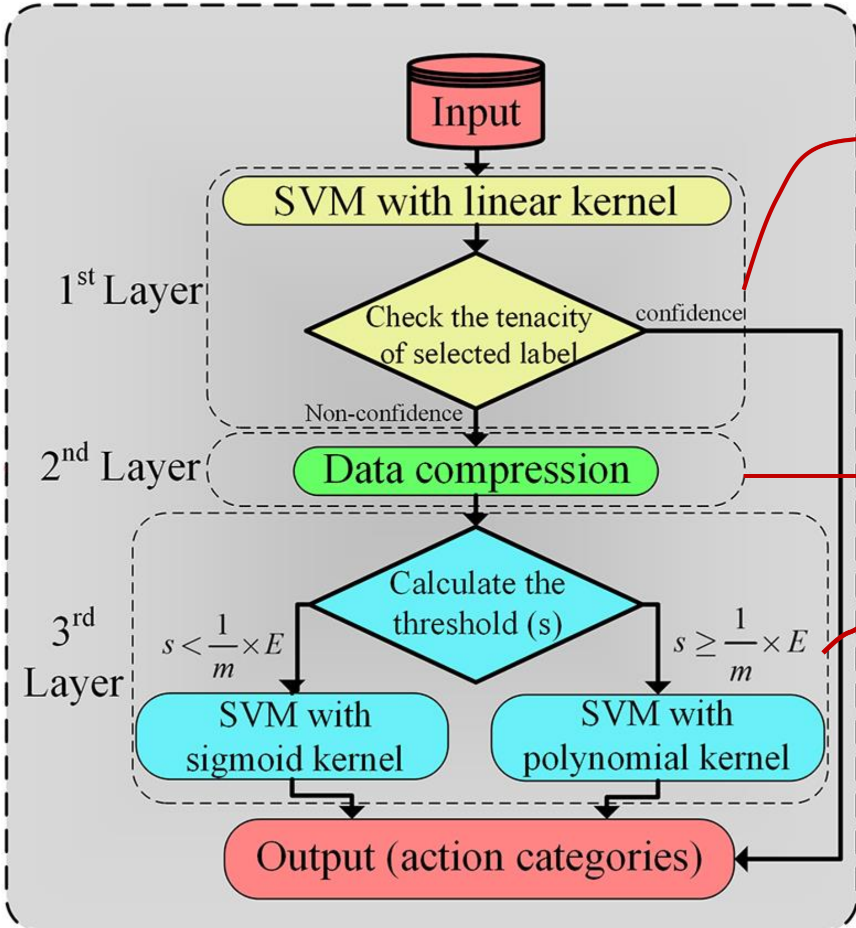
- We hypothesize that only the motion features can provide enough information to recognize actions.
- The Histogram of Optical Flow (HOF) along with Dense Trajectory features are utilized for feature extraction.

Fisher Vector Encoding (FV)

- FV requires Gaussian mixture models (GMMs) to build the vocabulary.
- We train a ***64*** component GMM to learn the $\lambda = \{\boldsymbol{\omega_k}, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\}|_{\boldsymbol{k=1}}^{K}$ parameters over a random subset of the training features.
- Given a video with the set of descriptors $(x_{1,\dots,}x_n)$,the FV becomes the concatenation of the normalized partial derivatives of means and deviations

University of Windsor

Hybrid Classifier



1st Layer

- **Thresholding Calculation:** $\tau = \max_1(P) - \max_2(P) + \frac{1}{m}$
- **In case of providing** $\tau \leq \max_1(P)$ **the result of linear SVM is considered as non-confident and the features are passed to the second layer.**

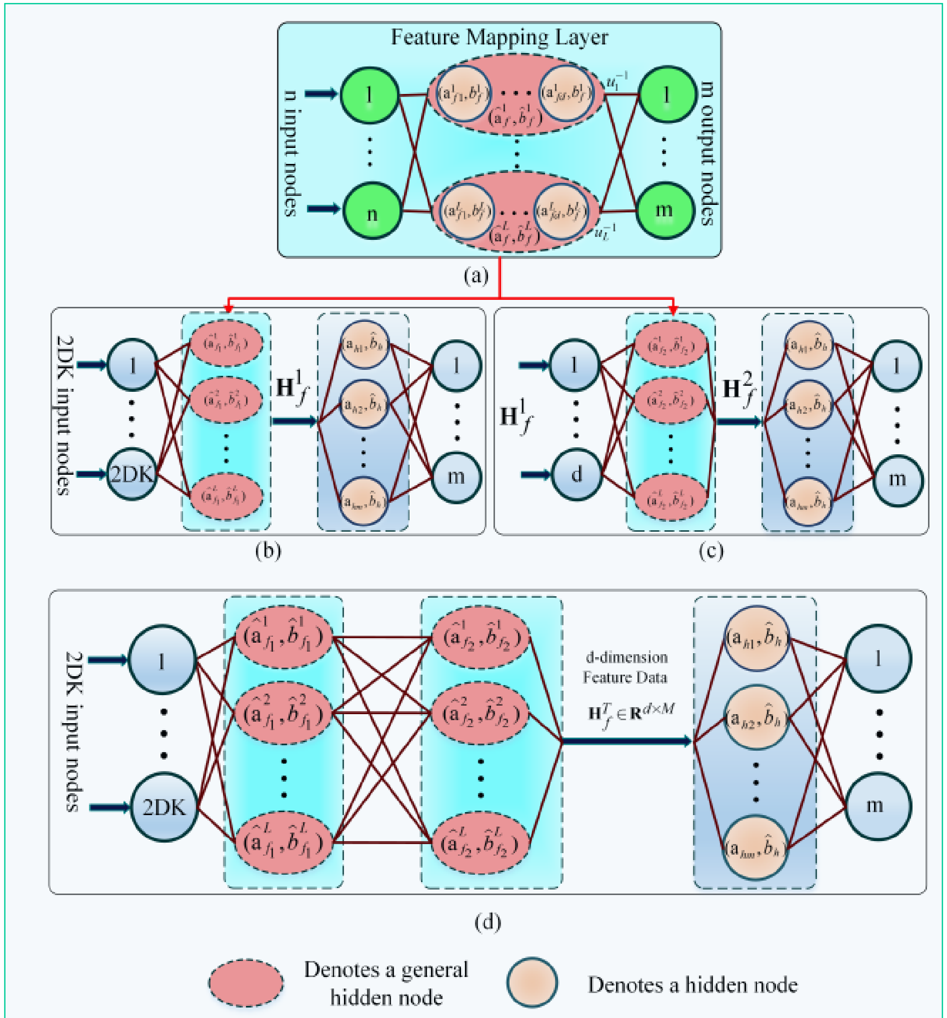
2nd Layer

- **Compress the encoded features to d dimension.**
- **We employ the double layer net with sub-network nodes to efficiently extract the most informative data from the encoded features.**

3rd Layer

- **The experiments demonstrate that SVM with sigmoid and polynomial kernels obtain different recognition performances based on the compressed features.**
- $Sigmoid: \quad \max_1(P) - \max_2(P) < \frac{1}{m} \times E$
  $Polynomial: \quad \max_1(P) - \max_2(P) \geq \frac{1}{m} \times E$

Data Compression



a) demonstrates the feature mapping layer.

b) shows the first network for compressing the original data.

c) shows the second network for compressing the original data.

d) shows the combination of the first and second stages in the multi-layer network including two feature mapping layers.

Data Compression

The following steps are performed for data compression:

1) Randomly generate the initial general node of the feature mapping layer, by setting j = 1,

$$\mathbf{H}_f^j = \mathbf{g}\left(\hat{\mathbf{a}}_f^j . \mathbf{x} + \hat{\mathrm{b}}_f^j\right), \left(\hat{\mathbf{a}}_f^j\right)^T . \hat{\mathbf{a}}_f^j = \mathbf{I}, \left(\hat{\mathrm{b}}_f^j\right)^T . \hat{\mathrm{b}}_f^j = 1$$

where $\hat{\mathbf{a}}_f^j \in \mathbf{R}^{d \times 2DK}$, $\hat{\mathrm{b}}_f^j \in \mathbf{R}$.

2) Calculate the parameters in the learning layer based on the sigmoid activation function (g) for any continuous desired outputs (y),

$$\hat{\mathbf{a}}_h = \mathbf{g}^{-1}(u_{2DK}(\mathbf{y})) \cdot \left(\mathbf{H}_f^j\right)^{-1} \quad , \hat{\mathbf{a}}_h^j \in \mathbf{R}^{d \times m}$$

$$\hat{b}_h = \sqrt{\mathrm{mse}\left(\hat{\mathbf{a}}_h^j \cdot \mathbf{H}_f^j - \mathbf{g}^{-1}(u_{2DK}(\mathbf{y}))\right)} \, , \hat{b}_{2DK}^j \in \mathbf{R}$$

$$\mathbf{g}^{-1}(\cdot) = -\log\left(\frac{1}{(\cdot)} - 1\right) \quad \text{if } \mathbf{g}(\cdot) = 1/(1 + e^{-(\cdot)})$$

Where $\mathbf{H}^{-1} = \mathbf{H}^T\left(\frac{C}{\mathbf{I}} + \mathbf{H}\mathbf{H}^T\right)^{-1}$.

Data Compression

3) Update the output error:

$$\mathbf{e}_j = \mathbf{y} - u_{2DK}^{-1}\mathbf{g}(\mathbf{H}_f^j, \hat{\mathbf{a}}_h, \hat{b}_h)$$

4) obtain the error feedback data:

$$\mathbf{P}_j = \mathbf{g}^{-1}(u_{2DK}(\mathbf{e}_j)) \cdot (\hat{\mathbf{a}}_h)^{-1}$$

5) Update the feature data as $\mathbf{H}_f^j = \sum_{l=1}^{j} u_l^{-1}\mathbf{g}(\mathbf{x}, \hat{\mathbf{a}}_f^l, \hat{b}_f^l)$ by setting j = j +1 and adding a new general node $\hat{\mathbf{a}}_f^j, \hat{b}_f^j$:

$$\hat{\mathbf{a}}_f^j = \mathbf{g}^{-1}(u_j(\mathbf{P}_{j-1})) \cdot \mathbf{x}^{-1} \quad , \hat{\mathbf{a}}_f^j \in \mathbf{R}^{d \times 2DK}$$

$$\hat{b}_f^j = \sqrt{\mathrm{mse}(\hat{\mathbf{a}}_f^j \cdot \mathbf{x} - \mathbf{P}_{j-1})} \quad , \hat{b}_f^j \in \mathbf{R}$$

6) Repeat steps 2 to 4 for L-1 times. So, the optimal informative data are obtained by:

$$\mathbf{H}_f^L = \sum_{j=1}^{L} u_j^{-1}\mathbf{g}(\mathbf{x}, \hat{\mathbf{a}}_f^j, \hat{b}_f^j) = \mathbf{H}_f^*$$

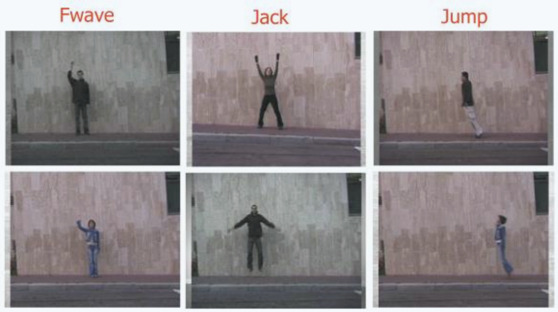
Data Compression

- The data compression can be used as a multi-layer network.
- The multilayer network provides a better general performance than single layer structure.
- In the multi-layer strategy, the input data is transformed into multi-layers, and the input encoded features is converted into d-dimensional space using multitude feature mapping layers.

- Thus, given a training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{M} \subset \mathbf{R}^{2DK} \times \mathbf{R}^{m}$, the compressed features are represented as $\mathbf{H}_f^T = \sum_{i=1}^{L} \mathbf{g}(\mathbf{H}_f^T \cdot \hat{\mathbf{a}}_f^i + \hat{b}_f^i)$ where $\mathbf{H}_f^T$ is the output of the second layer in the multi-layer network.

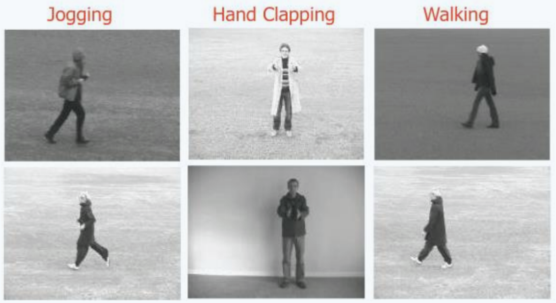
Datasets

1) **Weizmann dataset** contains 90 videos and 10 classes of simple actions. The evaluation of Weizmann is performed by leave one out cross validation.



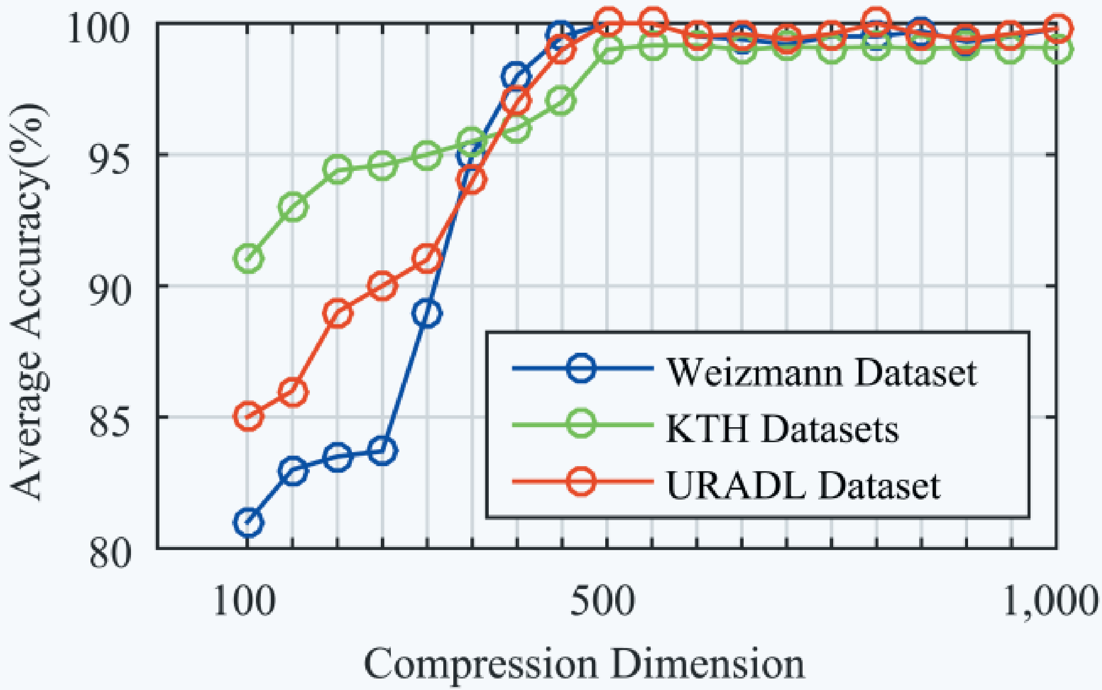
2) **URADL dataset** is a high resolution dataset of 10 complicated actions in 150 videos. The 10-fold cross validation is employed to evaluate this dataset.

3) **KTH dataset** contains six types of human actions. The evaluation of KTH dataset is performed based on 192
training and 216 testing samples.



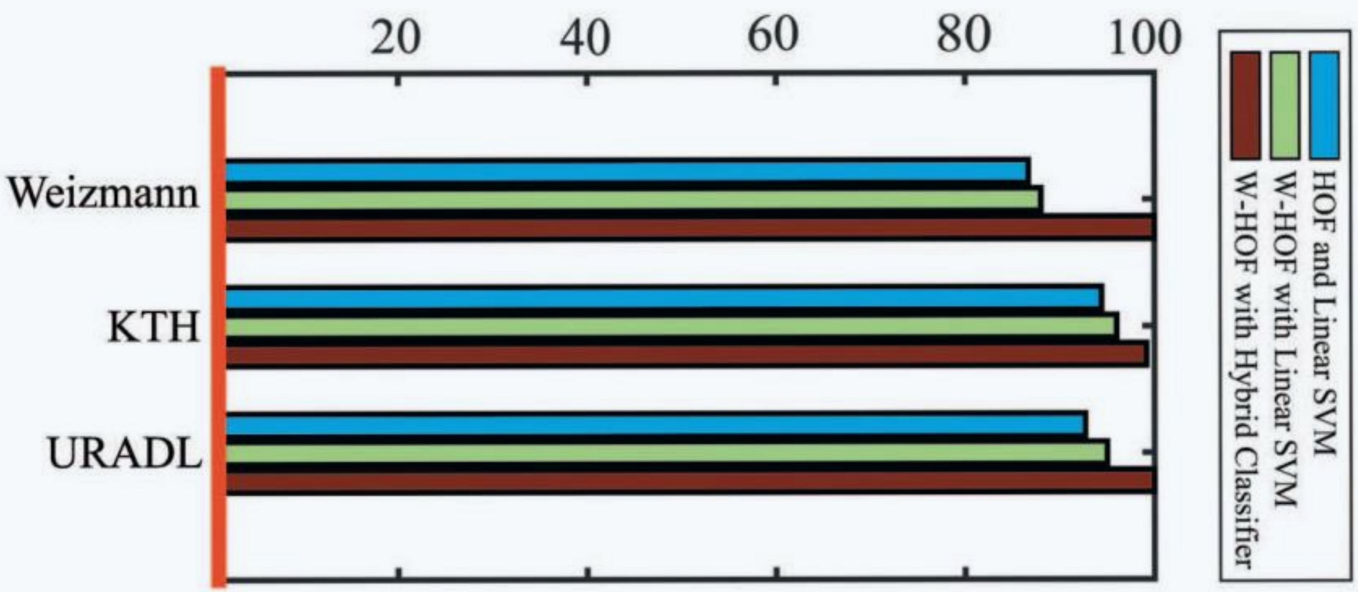
Experimental Results

Evaluation of a set of dimensions for compressing the features at the second layer of hybrid classifier.



Experimental Results

Simple action recognition performance using preprocessing steps and hybrid classifier.



Comparison with the state-of-the-arts

Dataset	Method	Recognition Rate
Weizmann	Cao et al. [23]	99.6%
	Lei et al. [24]	89.2%
	Samanta et al. [25]	90.0%
	Sushma et al. [26]	95.55
	Proposed Framework	**100.00%**
KTH	Cao et al. [23]	92.0%
	Lei et al. [24]	93.97%
	Samanta et al. [25]	94.7%
	Barrett et al. [27]	94.9%
	Proposed Framework	**98.00%**
URADL	Prest et al. [28]	92%
	Bilibski et al. [29]	94.7%
	Wang et al. [7]	96%
	Eman et al. [34]	96.6%
	Proposed Framework	**100.00%**

Conclusion

- We have Modified the Bag of Visual Word Framework for the simple action recognition by enhancing the following steps:

1. Propose the novel hybrid classifier to leverage the most informative parts of encoded features.
2. Evaluate the effect of using different SVM kernels on the compressed features.
3. Evaluate the effect of 3D Wavelet Transform as the preprocessing step for local feature extraction.

References

[1] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 12, pp. 2247–2253, 2007.

[2] M. Ross, P. Chris, and K. Henry, “Activity recognition using the velocity histories of tracked keypoints,” in Proc. ICCV, 2009.

[3] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local svm approach,” in Proc. ICPR, 2004, pp. 32–36.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Proc. CVPR, 2008, pp. 1–8.

[5] J. Zhu, B. Wang, X. Yang, W. Zhang, and Z. Tu, “Action recognition with actons,” in *Proc. ICCV*, Dec. 2013, pp. 3559–3566.

[6] K. Hildegard, J. Hueihan, G. Estbaliz, P. Tomaso, and T. Serre, “Hmdb: a large video database for human motion recognition,” in Proc. ICCV, 2011, pp. 2556–2563.

[7] H. Wang, A. Klser, C. Schmid, and L. Cheng-Lin, “Action recognition by dense trajectories,” in Proc. CVPR, Jun. 2011, pp. 3169–3176.

[8] L. Wang, Y. Qiao, and X. Tang, “Latent hierarchical model of temporal structure for complex activity classification,” *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 810–822, Feb. 2014.

[9] X. Peng, C. Zou, Y. Qiao, and Q. Peng, “Action recognition with stacked fisher vectors,” in *Proc. ECCV*, Sep. 2014, pp. 581–595.

[10] M. Sapienza, F. Cuzzolin, and P. H.S. Torr, “Learning discriminative space time action parts from weakly labelled videos,” *Int. J. Comput. Vis.*, vol. 110, no. 1, pp. 30–47, Oct. 2014.

[11] A. Gaidon, Z. Harchaoui, and C. Schmid, “Activity representation with motion hierarchies,” *Inter. jour. of comp. vis.*, vol. 107, no. 3, pp. 219–238, 2014

[12] X. Peng, L. Wang, X. Wang, and Y. Qiao, “Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice,” 2014. [Online]. Available: Available: http://arxiv.org/abs/1405.4506

[13] Y.-G. Jiang, Q. Dai, W. Liu, X. Xue, and C.-W. Ngo, “Human action recognition in unconstrained videos by explicit motion modeling,” *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3781–3795, 2015.

[14] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, “Rank pooling for action recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 4, pp. 773–787, 2016.

[15] P. Bilinski and F. Bremond, “Video covariance matrix logarithm for human action recognition in videos,” in *Proc. IJCAI*, Jul. 2015.

[16] L. Wang, Y. Qiao, and X. Tang, “Action recognition with trajectory pooled deep-convolutional descriptors,” in *Proc. CVPR*, Jun. 2015.

[17] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, “Modeling video evolution for action recognition,” in Proc. CVPR, 2015, pp. 5378–5387

[18] Wang, A. Klaser, C. Schmid, and C.-L. Liu, “Dense trajectories and motion boundary descriptors for action recognition,” International journal of computer vision, vol. 103, no. 1, pp. 60–79, 2013.

[19] L. Liu, L. Shao, X. Li, and K. Lu, “Learning spatio-temporal representations for action recognition: A genetic programming approach,” IEEE trans. on cyb., vol. 46, no. 1, pp. 158–170, 2016.

[20] F. Liu, X. Xu, S. Qiu, C. Qing, and D. Tao, “Simple to complex transfer learning for action recognition,” *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 949–960, Feb. 2016.

[21] Y. Wang, J. Song, L. Wang, L. Van Gool, and O. Hilliges, “Temporal segment networks: Towards good practices for deep action recognition,” in *Proc. ECCV*, Oct. 2016.

[22] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, “Hierarchical clustering multi-task learning for joint human action grouping and recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 266–278, Feb. 2016.

[23] X.-Q. Cao and Z.-Q. Liu, “Type-2 fuzzy topic models for human action recognition,” IEEE Trans. Fuzzy Sys., vol. 23, no. 5,pp. 1581–1593, 2015.

[24] J. Lei, G. Li, J. Zhang, Q. Guo, and D. Tu, “Continuous action segmentation and recognition using hybrid convolutional neural network-hidden markov model model,” IET Comp. Vis.,2016.

[25] S. Samanta and B. Chanda, “Space-time facet model for human activity classification,” IEEE Trans. Multimedia, vol. 16, no. 6, pp. 1525–1535, 2014.

[26] S. Bomma and N. M. Robertson, “Joint classification of actions with matrix completion,” in Proc. ICIP, 2015, pp. 2766–2770.

[27] D. P. Barrett and J. M. Siskind, “Action recognition by timeseries of retinotopic appearance and motion features,” IEEE Trans. Circ. Sys. Vid. Tech., vol. 26, no. 12, pp. 2250–2263, 2016.

[28] A. Prest, V. Ferrari, and C. Schmid, “Explicit modeling of human-object interactions in realistic videos,” Pat. Recog., vol. 35, no. 4, pp. 835–848, Apr. 2013.

[29] P. Bilinski and F. Bremond, “Video covariance matrix logarithm for human action recognition in videos,” in Proc. IJCAI, Jul. 2015, pp. 2140–2147.

[30] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, “Advances in human action recognition: A survey,” arXiv preprint arXiv:1501.05964, 2015.

[31] Y. Yang and Q. M. J. Wu, “Multilayer extreme learning machine with subnetwork nodes for representation learning,” IEEE Trans. Cybern., vol. 46, no. 11, pp. 2570–2583, Nov.2016

[32] M. Jain, H. Jegou, and P. Bouthemy, “Better exploiting motion for better action recognition,” in *Proc. CVPR*, 2013, pp. 2555–2562.

[33] A. Prest, V. Ferrari, and C. Schmid, “Explicit modeling of human-object interactions in realistic videos,” Pattern Recognition, vol. 35, no. 4, pp.835–848, Apr. 2013

[34] E. Mohammadi, Q. M. J. Wu, and M. Saif, “Human activity recognition using an ensemble of support vector machines,” in Proc. HPCS, Jul. 2016, pp. 549–554.

[35] Y. Yang and Q. M. J. Wu, “Extreme learning machine with subnetwork hidden nodes for regression and classification,” IEEE Trans. Cybern., vol. PP, no. 99, pp. 1–14, Nov. 2015.

Thank You