# Beyond Gaussian Pyramid: Multi-skip Feature Stacking for Action Recognition

Zhenzhong Lan, Ming Lin, Xuanchong Li, Alexander G. Hauptmann, Bhiksha Raj
School of Computer Science, Carnegie Mellon University

| HMDB51 (MAcc. %) | | Hollywood2 (MAP %) | | UCF101(MAcc. %) | |
|---|---|---|---|---|---|
| Simonyan *et al.* [5] | 57.9 | Jain *et al.* [1] | 62.5 | Wang *et al.* [6] | 85.9 |
| Peng *et al.* [3] | 61.1 | Oneata *et al.* [2] | 63.3 | Simonyan *et al.* [5] | 87.6 |
| Peng *et al.* [4] | **66.8** | Wang *et al.* [7] | 64.3 | Peng *et al.* [3] | 87.9 |
| MIFS (L=3) | 65.1 | MIFS (L = 3) | **68.0** | MIFS (L = 3) | **89.1** |

Table 1: Comparison of our results to the state-of-the-arts.

This paper introduces a video feature enhancing technique called Multi-skIp Feature Stacking (MIFS), which stacks features extracted from videos with multiple frame rates. The only difference between MIFS and other conventional methods is that instead of only using feature extracted from one time scale, we extract and stack all the raw feature from different scales (by skipping frames) together before encoding. We prove that MIFS enhances the learnability of differential-based features exponentially. The resulting feature matrices from MIFS have a much smaller conditional numbers and variances than those from conventional methods, as shown in Figure 1. Experimental results, as shown in Table 1, 2, 4 show significantly improved performance on challenging action recognition and event detection tasks when applied MIFS to Improved Dense Trajectory (IDT) [7].

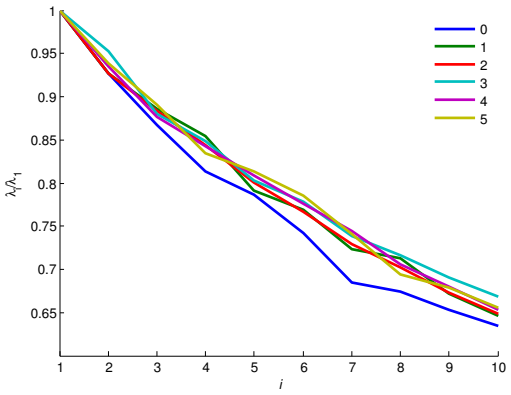


Figure 1: The decaying trends of singular values of feature matrices for UCF101 dataset. 0 to 5 indicate the MIFS level and $i$ indicates the $i$th singular value. We can see that MIFS representations do have a slower singular value decaying trend compared to conventional representations (blue lines).

Our baseline method is IDT by Wang & Schmid [7]. We follow all experimental setting as in [7] except we augment the descriptors of IDT with 3D normalized location information and perform another L2 normalization after concatenating the video-level descriptors of IDT.

Shown in Figure 1 are the trends of $\frac{\lambda_i}{\lambda_{max}}$ on the UCF101 dataset. As can be seen, the singular values of MIFS decrease slower than the conventional one (0). It is also interesting to see that by having one or two additional levels, we have already exploited most of the potential improvement.

Table 2 shows how performance changes with respect to the MIFS level. Our baseline performance on HMDB51, Hollywood2, and UCF101 datasets are 62.1% MAcc, 67.0% MAP and 87.3% MAcc respectively. These numbers are higher than Wang & Schmid [7]'s results, which are 57.2%, 64.3% and 85.9%, respectively. This is largely because our location extended descriptors and feature renormalization. From Table 2, we observe that for MIFS representations, they all perform better than single-scale representation and the performance decreasing points are later than those in the single-scale representations. We also observe that for MIFS representations, most of the performance improvement comes from $L = 1$ and $L = 2$.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

| | HMDB51 (MAcc%) | | Hollywood2 (MAP%) | | UCF101 (MAcc%) | |
|---|---|---|---|---|---|---|
| L | one-scale | MIFS | one-scale | MIFS | one-scale | MIFS |
| 0 | 62.1 | | 67.0 | | 87.3 | |
| 1 | 63.1 | 63.8 | 66.4 | 67.5 | 87.3 | 88.1 |
| 2 | 54.3 | 64.4 | 62.5 | 67.9 | 85.5 | 88.8 |
| 3 | 43.8 | 65.1 | 60.5 | **68.0** | 81.3 | **89.1** |
| 4 | 24.1 | **65.4** | 58.1 | 67.4 | 74.6 | **89.1** |
| 5 | 15.9 | **65.4** | 54.4 | 67.1 | 66.7 | 89.0 |

Table 2: Comparison of different scale levels for MIFS.

Table 1 shows that our method exceeds the state-of-the-arts on Hollywood2 and UCF101 and is comparable to state-of-the-arts on HMDB51 dataset.

MIFS can also be used as a speedup strategy for feature extraction with minimal or no accuracy cost. For example, removing L=0 (original videos) will significantly reduce cost but still give useful improvements as shown in Table 3. $L-1$ shows the results of only using features from every 2nd frame and $L = 2-0$ shows the results of combining features from level 1 (every 2nd frame) and level 2 (every 3rd frame) but not L=0. As can be seen, in most of cases, we can still get better results with less cost.

In Table 4, we also show that MIFS can also improve the performance of complex event detection, demonstrated by results from TRECIVD Multimedia Event Detection datasets.

| | HMDB51 (MAcc%) | Hollywood2 (MAP%) | UCF101 (MAcc%) | Computational Cost (Relative) |
|---|---|---|---|---|
| L=0 | 62.1 | 67.0 | 87.3 | 1.0 |
| L=1-0 | 63.1 | 66.4 | 87.3 | 0.5 |
| L=2-0 | 63.9 | 67.6 | 88.5 | 0.75 |

Table 3: Performance versus computational cost for feature extraction

| | MEDTEST13 | | MEDTEST14 | |
|---|---|---|---|---|
| | EK100 | EK10 | EK100 | EK10 |
| Baseline | 34.2 | 17.7 | 27.3 | 12.7 |
| MIFS (L=3) | **36.3** | **19.3** | **29.0** | **14.9** |

Table 4: Performance Comparison on the MED task.

[1] Mihir Jain, Hervé Jégou, and Patrick Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013.

[2] Dan Oneata, Jakob Verbeek, Cordelia Schmid, et al. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, 2013.

[3] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *arXiv preprint arXiv:1405.4506*, 2014.

[4] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng. Action recognition with stacked fisher vectors. In *Computer Vision–ECCV 2014*, pages 581–595. Springer, 2014.

[5] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[6] Heng Wang and Cordelia Schmid. Lear-inria submission for the thumos workshop. In *ICCV Workshop*, 2013.

[7] Heng Wang, Cordelia Schmid, et al. Action recognition with improved trajectories. In *ICCV*, 2013.