DA-VLAD: DISCRIMINATIVE ACTION VECTOR OF LOCALLY AGGREGATED DESCRIPTORS FOR ACTION RECOGNITION

Fiza Murtaza[1] Muhammad Haroon Yousaf MIEEE[1] Sergio A. Velastin SMIEEE[2,3,4]

[1] Univ. of Engg. & Tech. Taxila, Pakistan [2] Univ. Carlos III de Madrid, Spain [3] Cortexica Vision Systems Ltd., UK [4] Queen Mary Univ. of London, UK

Goal

- Recognize human actions in videos by utilizing discriminative power of action codewords.

Writing on board | Playing Cello | Cricket Shot

Fencing | Flicflac | Fall-floor

Motivation

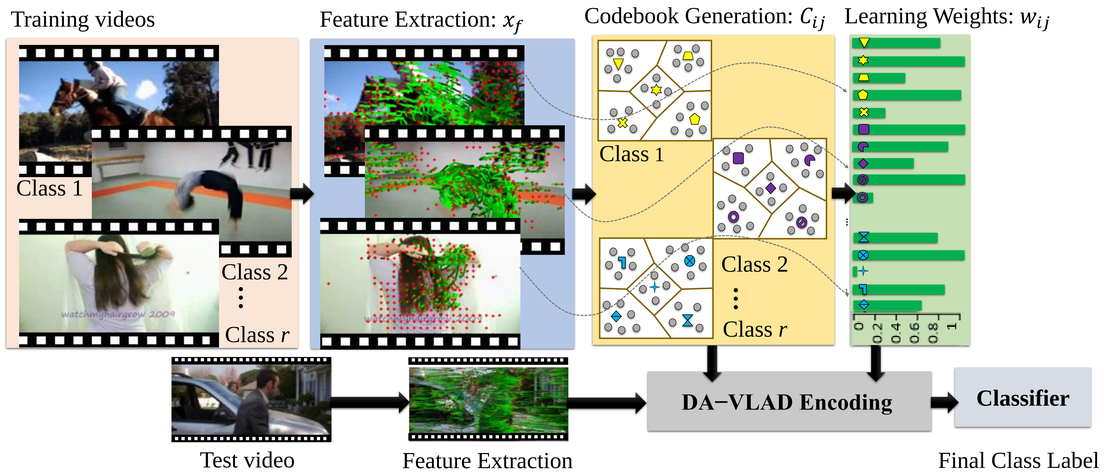
- Many overlapping frames between different actions cause non-discriminative codewords as shown below:

Discriminative power of the frames taken from HMDB51 dataset for action 'Hit' (top) and 'Punch' (bottom).

Contribution

- Propose a novel encoding approach, DA-VLAD which diminishes the effect of common codewords.
- Experiments show the importance of discriminative codewords for action recognition.
- DA-VLAD with Improved Dense Trajectories (IDT) improves the state-of-the-art results on UCF101 and HMDB51 dataset.

Approach



DA-VLAD encoding

$v_{ij} = w_{ij} \times \frac{1}{N_{ij}} \sum_{f=1}^{N_{ij}} (x_f - C_{ij})$, N_{ij}: number of features descriptors x_f assigned to C_{ij}

Final VLAD vector is formed by the concatenation of v_{ij}, $\forall i \in [1:r]$ and $j \in [1:K]$, where K is number of clusters and r is number of classes

Algorithm to learn Weights w_{ij}

Input: Feature descriptors $\{x_1, \dots, x_n\}$ and codewords C_{ij}

Output: q_{ij} and q'_{ij}

Procedure:

$q_{ij} = 0$, $q'_{ij} = 0$

for all actions $a_i \in A$ **do**

  for all features $x_f \in a_i$ **do**

    Assign x_f to the codeword C_{ij} such that $||x_f - C_{ij}||$ is minimum and find q_{ij} and q'_{ij}

    if $C_{ij} \in a_i$ **do**

      $q_{ij} = q_{ij} + 1$

    else

      $q'_{ij} = q'_{ij} + 1$

    end if

  end for

end for

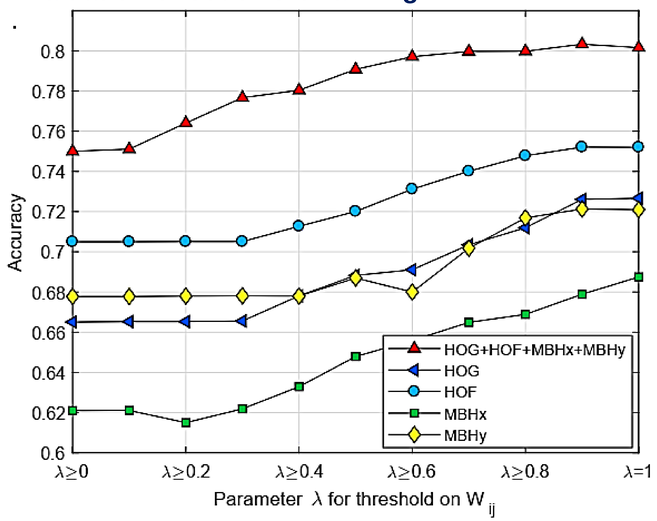
$w_{ij} = \frac{q_{ij}}{q_{ij} + q'_{ij}}$, $\forall i \in [1:r]$ and $j \in [1:K]$

Results

Datasets:

HMDB51: 6,766 video clips from 51 action categories. **UCF101:** 13,320 realistic video clips from 101 action classes.

Evaluation of codeword ranking



Comparison with other methods

Methods	UCF101	HMDB51
DT+MVSV [23]	83.5	55.9
iDT+iFV [6]	85.9	57.2
iDT+Hybrid [22]	87.9	61.1
iDT+MoFAP [19]	88.3	61.7
iDT+C3D [10]	90.4	-
iDT+C3D+ AdaScan [18]	93.2	66.9
iDT+GRP [20]	92.3	67.0
iDT+LTC [9]	92.7	67.2
iDT+ST-VLAD [16]	91.5	67.6
iDT+Two-Stream Fusion [8]	93.5	69.2
iDT+ActionVLAD(VGG-16) [17]	93.6	69.8
iDT+ST-VLMPF [15]	94.3	73.1
Our: iDT+DA-VLAD	**95.1**	**80.1**

- DA-VLAD outperforms the state-of-the-art approaches

Conclusion

- DA-VLAD exploited the discriminative power of action codewords, represented in the form of weights.
- These weights are integrated with standard VLAD encoding.
- Results show that action codewords with weight =1, i.e. λ=1, result in higher accuracy.
- In future, we can use DA-VLAD with CNN features.
- Discriminative weights can also be used with iFV and other encoding schemes.