Loss Switching Fusion with Similarity Search for Video Classification

Lei Wang*† Du Q. Huynh† Moussa Reda Mansour*†

*iCetana Pty Ltd †The University of Western Australia

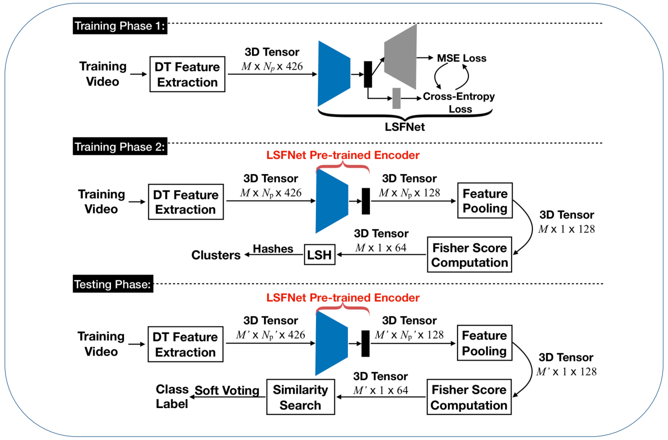
Introduction

- ❖We propose a novel video classification system that would benefit the scene understanding task.
- ❖Our classification problem: classifying background and foreground motions using the same feature representation for outdoor scenes.
- ❖We propose a lightweight Loss Switching Fusion Network (LSFNet) for the fusion of spatiotemporal descriptors and a similarity search scheme with soft voting to boost the classification performance.
- ❖Potential applications: content-based video clustering, video filtering, etc.

Method

➢Two training phases:

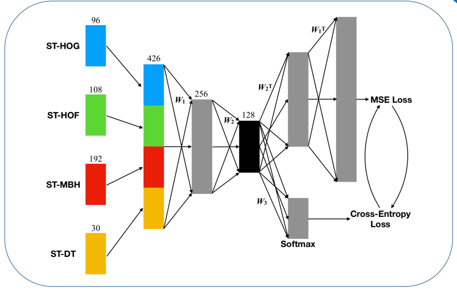
- LSFNet is trained using randomly sampled descriptors;
- The pre-trained LSFNet and a feature pooling layer together output a lower-dimensional feature vector.



LSFNet is composed of:

(i)a 5-layer autoencoder;

(ii)a multilayer perceptron (MLP) classifier shares the encoder part of the autoencoder.



➢The MSE loss and classification loss of LSFNet are used alternately in each pass of the gradient decent.

➢Locality Sensitive Hashing (LSH) is used to map features to a hash value.

➢For each test video, similarity search is used to find the most similar feature representations so as to get their corresponding labels.

➢Counting and comparing the number of labels retrieved using 'soft voting' to get the confidence values to assign label to each test video.

Datasets and Experimental Settings

➢Two industry datasets:

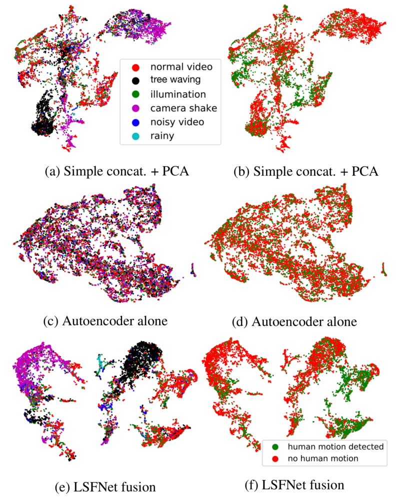
- **iCetanaPrivateDataset**
  - 2700 videos with various length captured in outdoor environments,
  - contains many background motions such as tree waving, camera shaking, rainy, noisy, etc.
- **iCetanaEventDataset**
  - An extension of iCetanaPrivateDataset
  - 6668 videos captured by multiple cameras located at different train stations, bus stops, etc.

➢Multi-class classification for 6 background motions;

➢Binary classification for separating human motions from background motions.

Experimental Results

- **Video Clustering**: Feature space visualization using UMAP for background and foreground motions on the testing set of the iCetanaPrivateDataset.



(a) Simple concat. + PCA (b) Simple concat. + PCA

(c) Autoencoder alone (d) Autoencoder alone

(e) LSFNet fusion (f) LSFNet fusion

- **Video Classification**: A comparison of our method with other state-of-the-art techniques.

Algorithms	Background env. motion	Foreground human motion
iDT [30]	48.1	66.7
C3D [20] (Sports 1M pre-training) + LinearSVM	74.1	70.4
C3D [20] (finetuned using iCetanaEventDataset)	75.9	77.8
I3D RGB [21](finetuned using iCetanaEventDataset)	77.0	79.9
Fisher score + CCA†	81.5	85.2
DT + FV + Fisher score + LSH‡	83.8	86.5
LSFNet	83.3	85.2
LSFNet+ Fisher score	85.2	87.0
Our whole system	**88.9**	**90.7**

†Our own pipeline using Fisher score for each spatiotemporal descriptor followed by Canonical Correlation Analysis (CCA) [3] for the feature fusion.

‡Our own pipeline using DT [26] followed by Fisher vector (FV) [37, 38], then Fisher score is used to select the top-50% feature components for LSH.