# FEATURES FOR VIDEO CLASSIFICATION

**Kamil Behúň**
Master Degree Programme (2), FIT BUT
E-mail: xbehun03@stud.fit.vutbr.cz

Supervised by: Michal Hradiš
E-mail: ihradis@fit.vutbr.cz

**Abstract**: This work compares hand-designed features with features learned by Independent subspace analysis (ISA) in video classification. The features learned by ISA were tested in a standard Bag of Visual Word classification paradigm replacing hand-designed features (e.g. SIFT, HOG, C2). The classification performance was measured on Human Motion DataBase where they show supperior performance over the hand-desined features.

**Keywords**: HMBD, ISA, video classification, feature extraction

## 1 INTRODUCTION

In recent years, video production increased signifficantly. These videos include movies, series, TV shows, etc. But main part of these videos is created by amateurs, because nowdays anyone can create a video. This production rise a need to organize the data for efficient manipulation, access and search. Video classification can be used for this task [2, 3].

This work is primarily aimed at the first fundamental stage of video classification – feature extraction. Speciffically to compare local descriptors learned from the data with commonly used hand-designed descriptors like SIFT, HOG/HOF.

## 2 PREVIOUS WORK

Classic approach to video classification can be divided into two main parts. These parts are feature extraction and classification.

There are two main information channels in video. The first one is the sound channel, which contains sounds from environment and the second one is visual channel with visual information. Most methods use the visual channel for feature extraction, because the visual channel contains richer information: shape of object, changes of objects, movement, etc.

Visual features can be divided into two groups [2]. The first group contains features based on frames, which are extracted from each frame independently. Frames are chosen randomly, sequentially or by a clustering method. The chosen frames can be described by global features (whole frame is described by feature) or local features (part of frame is described by feature). The most popular global feature is the GIST descriptor [6], which is a low-dimensional representation of a scene and it represents the dominant spatial structure of the scene. Local features are extracted from stable patches or from patches which are defined by a grid. The main local descriptor is SIFT descriptor [2], which represents gradients in a local area, and which is partially invariant to scale, rotation and small affine transformations.

The second group of visual features is group of spatio-temporal features, which are extracted from spatio-temporal video volumes. There are two types of spatio-temporal features: Spatio-temporal local features and trajectory features. Spatio-temporal local features expand local features based on

frames into 3D. Examples of these features are Cuboid, 3D-SIFT, HOF/HOF, C2 shape features, etc [2, 7]. HOF/HOG is combination of histogram of oriented gradient HOG (captures appearance) and histogram of oriented optical flow HOF (captures a movement). Trajectory features describe movement of local features in video frames. An example of trajectory features is a motion boundary histogram (MBH), which computes derivates of optical flow [2].

The local features are usually aggregated into Bag of Visual Words (BOW) representation to create compact representation of the whole video [2]. This BOW representation is next used as input to classifier.

Any classifier can be used. However, Support Vector Machine (SVM) is the most common [3, 5, 4].

## 3 INDEPENDENT SUBSPACE ANALYSIS

Independent subspace analysis (ISA) is a unsupervised learning algorithm [5]. ISA can be described as a two layer network with a square activation function in the first layer and square-root activation function in the second layer. Each of the second layer units pools over a small number of units of the first layer. Weights in first layer are trained and weights in the second layer are fixed. The weights in the second layer represent structure of the subspace of neurons in the first layer. ISA is learned by changing values of weights in the first layer targeting to mizimize output of the second layer over training data, while weights of the neurons in the first layer are forced to be ortogonal. This orthogonality ensures diversity of the output features.

The output of ISA network can be used as a new efficient representation of the original data. The obtained features are robust to translation and selective to frequency and rotation changes, which is good for classification [5].

## 4 METHOD

To compare local ISA descriptors learned from a data with commonly used hand-designed descriptors, we created a classification system with two feature extraction methods.

The first of these methods extracts features from small local spatio-temporal video volumes ($16 \times 16$ pixels in 11 frames). Volumes are selected by uniform grid from video. The selected volumes are linearized to a vector. A dimension of this vector is reduced by PCA and result is described by ISA.

The second method is used as a baseline. This method extracts features from sequence of frames, which are selected from video at regular intervals. Each frame is described by several SIFT descriptors independently.

To create a feature vector for a whole video, the extracted descriptors by both methods are aggregated to the BOW representation [1]. In order to obtain the BOW representation, descriptors are at first translated to visual words by codebook transform. K-means algorithm with Euclidean distance is used to obtain the set of prototypes which constitutes the codebook – cluster centers become the prototypes. In the experiments, 4096 codewords are used. A separate codebook is created for each method.

Experiments were performed with Support Vector Machine classifier (SVM) and $\chi^2$ kernel.

$$K(H_i, H_j) = exp\left(-\gamma \sum_{n=1}^{V} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}\right), \tag{1}$$

where $H_i = \{h_{in}\}$ and $H_j = \{jn\}$ are BOW representations and $V$ is size of BOW. Optimal value of the SVM regularization parameter $C$ and the Gaussian kernel scale $\gamma$ were estimated by grid search with 6-fold cross-validation with stratified sampling of training dataset.

| features + classifier | Classification accuracy |
|---|---|
| C2 + SVM [4] | 22.83 % |
| ISA + SVM (this work) | 22.03 % |
| HOG/HOF + SVM [4] | 20.44 % |
| SIFT + SVM (this work) | 19.26 % |

**Table 1:** Classification accuracies achieved by our approach and published approaches for HMDB.

## 5 EXPERIMENTS AND RESULTS

The system was tested on Human Motion DataBase (HMDB). This Dataset contains 51 motion classes and is splitted into training (3570 video clips) and testing part (1350 video clips). This dataset is free and is described in [4]. Table 1 shows the result complemented by prevous result by Kuehne et al. for HMDB dataset. Kuehne et al. used hand-designed feature extraction methods and SVM with $\chi^2$ kernel. We did the same experiment Kuehne et al. described in [4].

As can be seen in Table 1, the proposed ISA approach provides superior performance compared to other published results with hand-designed feature learning methods for HMDB in [4]. The table shows the best result which was achieved by ISA with PCA reduction to 330 dimensions and 2 subspace size. As can be seen in this table, ISA achieved better result than SIFT in our experiments.

## 6 CONCLUSION

The experiments show that the proposed ISA feature extraction provides state-of-the-art results in video classification. Specifically, its performance is superior to previously published results on the HBDB dataset and our results with SIFT. We plan to test other methods of feature learning in video classification like Autoencoders and Restricted Boltzmann Machines.

## REFERENCES


[1] M. Hradiš, I. Řezníček, and K. Behúň. Semantic class detectors in video genre recognition. In *Proceedings of VISAPP 2012*, pages 640–646, 2012.

[2] Y. Jiang, S. Bhattacharya, S. Chang, and M. Shah. High-level event recognition in unconstrained videos. In *International Journal of Multimedia Information Retrieval*, pages 1–29, 2012.

[3] Y. Jiang, Ch. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 494–501. ACM, 2007.

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2556–2563, 2011.

[5] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3361–3368, 2011.

[6] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, pages 145–175, 2001.

[7] M. Sapienza, F. Cuzzolin, and P. Torr. Learning discriminative space-time actions from weakly labelled videos. pages 123.1–123.12. BMVA Press, 2012.