Noname manuscript No.
(will be inserted by the editor)

A Modified Vector of Locally Aggregated Descriptors Approach for Fast Video Classification

Ionuţ Mironică · Ionuţ Cosmin Duţă · Bogdan Ionescu · Nicu Sebe

Received: date / Accepted: date

Abstract In order to reduce the computational complexity, most of the video classification approaches represent video data at frame level. In this paper we investigate a novel perspective that combines frame features to create a global descriptor. The main contributions are: (i) a fast algorithm to densely extract global frame features which are easier and faster to compute than spatio-temporal local features; (ii) replacing the traditional k-means visual vocabulary from Bag-of-Words with a Random Forest approach allowing a significant speedup; (iii) the use of a modified Vector of Locally Aggregated Descriptor(VLAD) combined with a Fisher kernel approach that replace the classic Bag-of-Words approach, allowing us to achieve high accuracy. By doing so, the proposed approach combines the frame-based features effectively capturing video content variation in time. We show that our framework is highly general and is not dependent on a particular type of descriptors. Experiments performed on four different scenarios: movie genre classification, human action recognition, daily activity recognition and violence scene classification, show the superiority of the proposed approach compared to the state of the art.

Keywords Capturing content variation in time in video · Modified Vector of Locally Aggregated Descriptor · Random Forests · Video classification

I. Mironică, B. Ionescu
LAPI, University Politehnica of Bucharest, 061071 Romania
E-mail: {imironica, bionescu}@imag.pub.ro

I. C. Duţă, N. Sebe
DISI, University of Trento, Italy
E-mail: {duta, sebe}@disi.unitn.it

1 Introduction

Along with the advances in multimedia and Internet technology, a huge amount of data, including digital video and audio, are generated on daily basis. This makes video in particular one of the most challenging data to process. Video processing and analysis has been the subject of a vast amount of research in the information retrieval literature [1–4]. Until recently, the best video search approaches were mostly restricted to text-based solutions which process keyword queries against text tokens associated with the video, such as speech transcripts, closed captions, social data, and so on. Their main drawback is in the limited automation because they require human input. The use of other modalities, such as visual and audio has been shown to improve the retrieval performance [5], attempting to bridge further the inherent gap between the real world data and its computer representation. The target is to allow automatic descriptors to reach a higher semantic level of description, similar to the one provided by manually obtained text descriptors.

Existing state-of-the-art algorithms for video classification can achieve promising performance in benchmarking for many research challenges, starting from genre classification to event and human activity recognition [6–9]. Even if these methods are designed to solve a single application, they can be adapted to a broad category of video classification tasks. A weak point of many video processing approaches is in the content description and training frameworks, which stand as a basis for any higher level processing steps. For instance, deep learning techniques come with very high complexity which translates into significant processing time for optimizing the complex architectures of the networks. Also, video information is temporal data and one of its definitive information is given by the changing/moving content. There are a number of approaches that attempt in particular to capture that temporal information, e.g., local motion features, dense trajectories, spatio-temporal volumes [6] to provide better representative power. However, these approaches generate a large amount of data which may trigger a high computational complexity for large-scale video datasets. In order to reduce this amount of information, one lead is to exploit frame-based features, where each global feature captures information of a single video frame.

In this context, this article proposes a new video content description framework that is general enough to address a broad category of video classification problems while remaining computational efficient. It combines the fast representation provided by Random Forests and VLAD framework with the high accuracy and the ability of Fisher Kernels to capture temporal variations.

The remainder of the paper is organized as follows. In Section 2 we present the current state-of-the-art and situates our approach accordingly. Section 3 presents the framework and the implementation details. The experimental validation setup is presented in Section 4 while the experimental results are presented and discussed in Section 5. Finally, Section 6 concludes the paper and discusses future perspectives.

2 Related Work

Video content classification remains one of the most challenging video processing problems, mainly because it implies the classification of complex semantic categories from a huge volume of multimodal data. There are several major approaches that emerged in the last decade.

A large family of video classification methods is based on creating global descriptors by aggregating local spatio-temporal features. In this context, a standard video classification system consists of detecting sparse spatio-temporal interest points which are then described using local spatio-temporal features, e.g., Histograms of Oriented Gradients (HoG), Histograms of optical Flow (HoF), Motion Boundary Histograms (MBH) [6]. The features are then encoded into the Bag-of-Words (BoW) [11] representation and combined with a classifier. Spatio-temporal interest point-based methods represent the scene and the performed actions as a combination of local descriptors, which are computed in a neighborhood of some interest points. The neighborhood can be selected as an image patch or as a spatio-temporal volume, e.g., cuboid. The spatio-temporal interest point based methods have received a lot of attention in the vision community due to their robustness to scale and viewpoint changes. Recently, Wang et al. [7] suggested the use of a set of dense trajectories, where the local video volume moves spatially through time. Additionally, they propose a new method for extracting the optical flow. They obtained good improvements over the HoG and HoF descriptors. Nevertheless, combining their dense trajectory descriptors with both HoG and HoF descriptors still gives significant improvements over dense trajectories alone.

Instead of computing local video features over spatio-temporal cuboids, state-of-the-art shallow video representations [8] make use of dense point trajectories. The local descriptor approaches are less sensitive to noise or occlusion. However, these approaches, often make use of the BoW model, requiring the quantization of large amounts of data. Even though the interest points and the features are computed locally, each sequence is represented by a global histogram, which does not carry any spatial or temporal information. The main improvements consist in changing the global descriptors from Bag-of-Words to other more accurate representations, namely the use of the Fisher vector encoding [10, 13] or Vector of Locally Aggregated Descriptors (VLAD) representation [28]. In [66] the authors proposed the VLAT model based on the aggregation of tensor products of local descriptors. Nakayama [67] presented an extension of VLAD that encodes the second-order statistics using local Gaussian metrics. Both methods obtained promising results for the image classification tasks. On the other side, Uijlings et al. [27] proposed the use of VLAD and Fisher kernels for video classification. They used several speedup approaches for densely sampled HoG and HoF descriptors and investigated the trade-off between accuracy and computational efficiency for the video representation using either a k-means or hierarchical k-means based visual vocabulary, a Random Forests (RF) based vocabulary or the Fisher encoding. Liu et al. [15] extracted a combination of motion and static features that are integrated into

a PageRank algorithm to prune the static features using motion cues as an alternative way to motion compensation. The hybrid use of motion and static features improved the performance of their approach.

Another perspective is to use human tracking algorithms to perform video content classification. For instance, Ikizler-Cinbis and Sclaroff [14] extracted multiple features on the human, objects and scene, and employed a multiple-instance learning framework for human action recognition of YouTube videos. Yang and Ramanan [42] proposed a method for articulated human detection and human pose estimation in videos based on a new representation of deformable part models. They detect small bounding boxes with a multi-scale HoG descriptor, instead of complete body limbs, making their work more efficient because it prevents the problem of double counting. The body part detector combined with the HoF features obtained good results on daily living activities [43]. However, these framework are adapted to a specific task and requires the use of motion compensation for foreground estimation and the detection and tracking of the human in the scene, generating a high computational cost. The accuracy of the algorithm is highly correlated with the performance of the human detector.

Convolutional Neural Networks (CNNs) have been shown to be an effective class of models for understanding image content, giving state-of-the-art results on image and video recognition, segmentation, detection and retrieval [16–18]. The key enabling factors behind these results are in the techniques for scaling up the networks to tens of millions of parameters and massive labeled data, that can support the learning process. Under these conditions, CNNs have been proved to learn powerful and interpretable video multimodal features. However, from a computational perspective, CNNs require extensively long periods of training time to effectively optimize the millions of parameters that compose the model. This difficulty is further compounded when extending the model to video classification. In [57], Karpathy at al. propose a convolutional neural framework in the context of the large-scale video classification. In order to improve the system's speed they create a multi-resolution architecture that uses two video streams. Input frames are fed into two separate processing streams: a context stream that models low-resolution image and a fovea stream that processes high-resolution center crop.

The main contribution of this paper is in the introduction of a new content representation pipeline which is designed specifically for video classification. The efficiency of the approach is demonstrated by the generality of the proposed framework in terms of applicability, and particularly it is successfully tested on four different classification scenarios. The algorithm incorporates a novel approach for frame-based features word assignment that uses a set of pruned Random Forests (RF) which are specially adapted for fast classification. We also propose a modified VLAD representation that allows achieving both, fast and high performance. The following secondary contributions are identified:

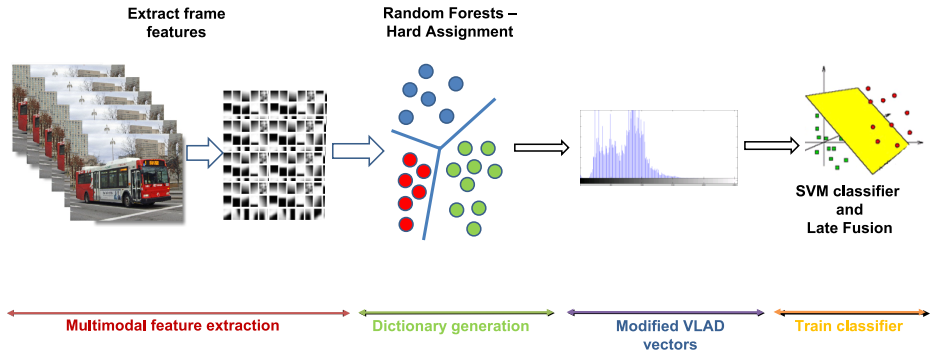


Fig. 1: Diagram of the proposed approach for video content representation.

(i) we introduce a modified Vector of Locally Aggregated Descriptor (VLAD) representation for video frame-based description that has the advantage of capturing content variation in time;
(ii) the proposed framework is feature independent in the sense that is not adapted to the use of a particular type of description scheme and can work basically with any content representation approach, from basic histograms, application dependent descriptors (e.g., body-part features) to more complex approaches that include feature points and multimodal integration;
(iii) we achieve similar or better performance than the state-of-the-art by using simpler and faster to compute descriptors.

3 Proposed Approach

The architecture of the proposed framework is presented in Figure 1. It consists of four processing steps:

- *step 1*: video data are represented with multimodal frame-based descriptors extracted with a dense sampling strategy;
- *step 2*: we create a dictionary of frame words using a new fast approach for word assignment. We replace the traditional k-means visual vocabulary from Bag-of-Words [6,11] with a Random Forest approach which allows for a significant speedup;
- *step 3*: each video frame is assigned to the nearest word (i.e., cluster center) using the previously trained random trees. Then, for each modality, we compute the proposed modified Vector of Locally Aggregated Descriptors (VLAD) [28];
- *step 4*: finally, the new resulting content descriptors are fed to a classifier which performs the actual classification task. We selected to use Support

Vector Machines (SVMs) due to their effectiveness in dealing with large feature vectors and sparse data. Also, this step acts as a late fusion mechanism by aggregating all individual modalities into a single decision.

Each of the processing steps is detailed in the following sections.

3.1 Feature extraction

In video retrieval, an important research problem is how to adequately capture temporal information. Some of the popular choices include the use of local cuboid features, e.g., motion boundary histograms or space-time interest points [19], but their main drawback is in the high computational complexity [20] (e.g., processing only few seconds of video may take tens of minutes). Another perspective is the use of global frame features, which are more computational effective [21, 22]. These approaches had obtained state-of-the-art results with low computational costs. In general, depending on the way the integration is carried out, frame based features may achieve similar or better performance than temporal approaches in many video classification problems. In our approach we use a set of multimodal frame-based features that are extracted using a dense sampling strategy. A detailed presentation of the experimented features is provided in Section 4.3.

3.2 Fast frame word assignment with Random Forests

Random Forests (RF) [23] are ensemble learning methods that operate by constructing a multitude of decision trees in the training stage and combine the classification output of all these trees. In our approach, we propose the use of RF for the word assignment step. There are several reasons that make RF a good candidate for this task, namely: (i) good classification accuracy proved in many scenarios [24, 25]; (ii) the computational time of RF word assignment is logarithmic with respect to the number of words due to the binary nature of the decision trees, whereas the computational time for the nearest neighbor approach is linear with respect to the number of visual words; (iii) RF trees are independent of the dimension of the descriptors, for splitting off each node only one dimension is selected and compared with a threshold, whereas the nearest neighbor approach is linearly dependent on the feature dimension.

Uijlings et al. [12] proposed several methods to speedup the Bag-of-Words classification pipeline for image classification using the RF for word assignment. They provided an evaluation of the efficiency of different visual word assignment strategies and showed that RF outperforms on computational time all the other approaches. Inspired by their work we propose the following improved approach.

We use a set of Random Forests that represent a combination of decision trees. Each tree is built independently using the same set of descriptors as input, e.g., $S = \{x_1, ..., x_n\}$, where n is the dimension. The leaves of a tree

represent the actual clusters. The construction of the trees is done in a recursive way. For each node, a number of split-offs are proposed by selecting at random one dimension of the descriptors. From these possible representations, we select only one that provides the maximum information gain (as defined in Equation 3). Each node splits the descriptors in two sets: S_L (left) and S_R (right). The process is repeated until a leaf is reached. Overall, each descriptor is fitted to a certain number of trees (i.e., a forest) with splits being selected from random choices. A common setting is to use hundreds of trees (a detailed parameter tuning is provided with the experimental results). This corresponds to the *training phase*. Similar approaches that use the information gain as a split scoring function are proposed in [64, 65]. They suggested that the improvement in predictive performance derived from improved information gain estimates, even small, is useful to many applications. However, these algorithms are presented in the context of classification and regression problems and not in particular as a word assignment strategy.

In the actual *clustering part*, descriptors are assigned to each (pre-trained) tree conducting to a certain path and a corresponding leaf (ending node). The video descriptor is then recomputed using this information (the process is detailed in Section 3.3). This leads to a content representation per tree. The final content description is achieved by simple concatenation of each individual tree representations.

For high accuracy, it is better to have a large number of trees as well as reaching a higher depth within the trees. However, increasing the number of trees and the depth results in a significant increase of the final descriptor dimension given by:

$$dim = 2^{depth} \times nTrees \times dimDesc \times 2 \tag{1}$$

where $depth$ represents the depth of the trees, $nTrees$ is the total number of trees for the Random Forests, $dimDesc$ is the initial descriptor dimension. The final multiplication by 2 results from the way we represent the modified VLAD vectors (details are presented in Section 3.3).

To have a numeric reference for the size of the output descriptors consider the case of 4 trees with a depth of 8 and an initial descriptor with 72 dimensions which results in a Random Forest descriptor of size 147,456. Increasing the number of initial trees from 4 to 10 and the depth to 10 we now obtain a final descriptor of 1,474,560 values, which is significantly higher. This large dimensionality can generate many computational problems, especially when targeting large scale classification scenarios. In particular, to perform some operations on this high dimensionality vectors, more memory resources are needed. Therefore, the computational time for manipulating these vectors grows significantly. Not least, we will demonstrate in Section 5.1 that reducing the size of final feature vector increases the classification performance for this approach.

To find a way for decreasing directly the dimensionality of the resulting feature vectors without any additional steps and maintaining the accuracy, we propose a novel approach that prunes the random forests. We use the

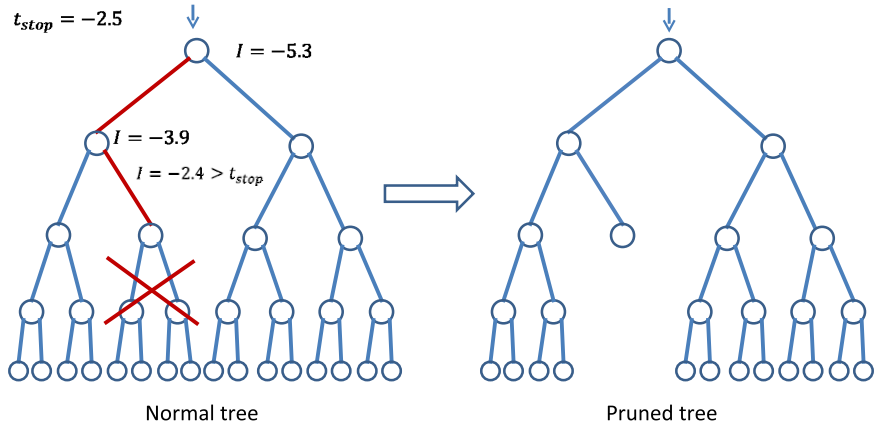


Fig. 2: Dimensionality reduction using information gain.

information gain of a node, I, for determining the best split-off point in the tree. It is defined as:

$$I = -\frac{\#(S_L)}{\#(S)} H(S_L) - \frac{\#(S_R)}{\#(S)} H(S_R) \tag{2}$$

where $\#(.)$ denotes the cardinality of the set and H is the Shannon Entropy of the class labels of the descriptors [30] given by:

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i) \tag{3}$$

where $P(x_i)$ represents the probability that a descriptor can reach the leaf i from the tree, and $H(S_L)$ and $H(S_R)$ represent the entropy for the left and right branches of the node. To stop the tree from growing, we use this measure to remove the branches that have an information gain lower than a certain threshold (t_{stop}). The process is depicted in Figure 2.

The intuitive idea behind this solution is that some nodes of the tree can contain descriptors which belong to the same class (or most of them belong to the same class). In this case we can state that the node is pure (or almost pure). Therefore, reaching a node of this kind will stop the split-off before reaching the given depth. The threshold for information gain is obtained empirically and represents a trade-off between the accuracy and the dimensionality of the resulting feature vector.

3.3 VLAD algorithm for frame-based words

A classic approach to compute the final feature vector is to take the frequency of the visual words and build a histogram. This histogram is fed to the final

Fig. 3: Different frame aggregation strategies: (a) simple frame mixing, (b) keyframe selection, (c) proposed approach (images from *Blip10000 [32]*).

classifier. When Random Forests are used for word assignment, the frequency of visual words represents the number of descriptors from each leaf. However, it was shown that this representation is outperformed by adopting Fisher Kernel [27] and VLAD representations [29].

A second novelty of our approach is in the way we compute the feature vector. The proposed approach exploits the advantages of both, Fisher Kernel and VLAD approaches, in a unified framework. Feature vectors are given by the concatenation of the $v_{\mu,i}$ and $v_{\sigma,i}$ Fisher representations for $i = 1, ..., K$ with K the number of words (number of clusters), given by:

$$v_{\mu,i} = \frac{1}{T\sqrt{P(x_i)}} \sum_{t=1}^{T} \frac{(x_t - \mu_i)}{\sigma_i} \tag{4}$$

$$v_{\sigma,i} = \frac{1}{T\sqrt{2P(x_i)}} \sum_{t=1}^{T} \left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right] \tag{5}$$

where x_t represents the frame-based features that are assigned to cluster t, μ_i is the mean of the training frame-based features for each cluster, σ_i is the standard deviation for cluster i, T is the number of descriptors from a cluster, $P(x_i)$ is the probability that a descriptor reaches a specific leaf from the tree.

To illustrate the benefits of this approach, a visual example is presented in Figure 3: Figure 3(a) presents a simple frame aggregation strategy (averaging) over the entire video sequence and Figure 3(b) presents the selection of a keyframe, which discards time information. In the proposed approach, interpreting the formulas in terms of variation in time, equation 4 averages the features over time, which are related as they fall in the same mixture component. Equation 5 models the variation of related features over the entire video sequence, capturing subtle visual changes. The different mixture components capture drastic variations in time such as shot changes specific to video.

This will result in a representation as presented in Figure 3(c), which captures better the content variation from different moments of time.

Globally, this approach can be interpreted as a hard assignment Fisher Kernel approach. The main differences between Fisher Kernel and our approach are the following: (i) we use a fast clustering method with Random Forests instead of the Gaussian Mixture Model (GMM); (ii) we perform a hard assignment strategy, rather than a soft assignment. We choose the hard assignment because it can be combined in an efficient way with the Random Forest word assignment approach. The contribution of the VLAD model is in the increase of the system's accuracy. By combining the Random Forests with the modified VLAD representations, we achieve both a fast and a high performance video classification system, which outperforms the other variants - as show in the experimental results (see Section 5).

3.4 Classification

The final component of the system consists of the data classifier which is fed with the descriptors issued for the proposed Fisher Kernel - VLAD approach. Among the broad choice of existing classification approaches [31] we selected a SVM classifier. We use several type of kernels, i.e., a fast linear kernel and two non-linear kernels: RBF and Chi-Square. While linear SVMs are very fast in both training and testing, SVMs with non-linear kernels are more accurate in many classification tasks due to better adaptation to the shape of the clusters in the feature space.

Finally, in the case of multimodal features, we combine the SVMs output confidence values using a linear weighted combination (late fusion):

$$CombMean(d,q) = \sum_{i=1}^{N} \alpha_i \cdot cv_i \tag{6}$$

where cv_i is the confidence value of classifier i for class q ($q \in \{1,...,C\}$), C represents the number of classes, d is the current video, α_i are some weights and N is the number of classifiers to be aggregated. The weights are learned during the optimization process that takes place on the training data, as presented in Section 5.

4 Experimental Setup

In this section we present the evaluation framework (dataset and metrics) and the choice of content descriptors.

4.1 Datasets

For testing the proposed video content description approach we have selected a broad range of classification scenarios, namely: video genre classification,

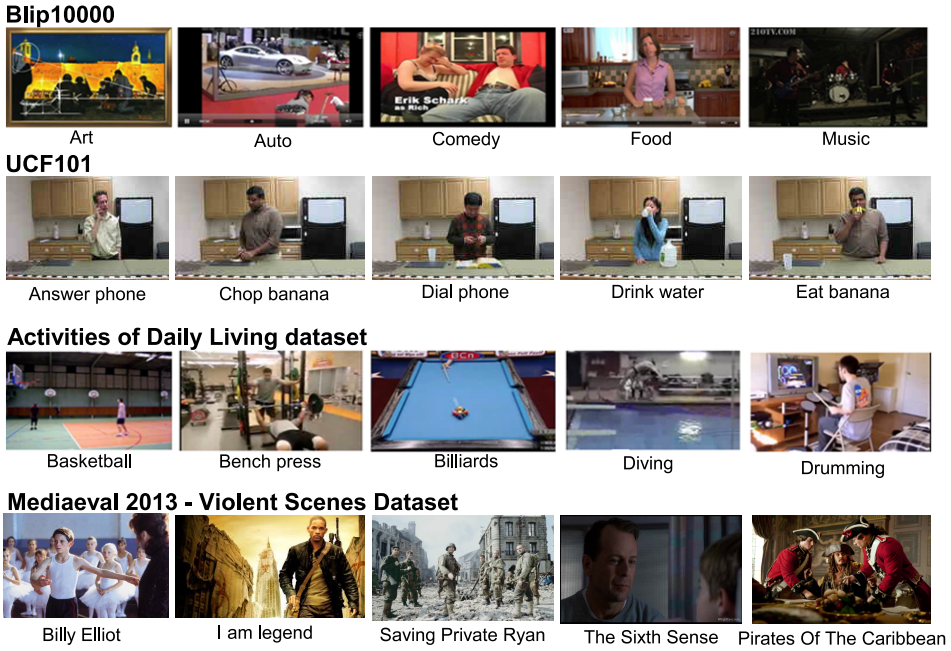


Fig. 4: Sample images from the experimentation datasets: Blip10000 [32], ADL [35], VSD2013 [36] and UCF101 [34].

classification of daily activities, classification of violent content and action recognition. We experimented on the following standard and publicly available datasets:

Blip10000 [32]: consists of 15,000 video sequences (around 3,250 hours of footage) retrieved from blip.tv[1]. Each video is labeled according to 26 web specific video genre categories: art, autos and vehicles, business, citizen journalism, comedy, conferences and other events, documentary, educational, food and drink, gaming, health, literature, movies and television, music and entertainment, personal or auto-biographical, politics, religion, school and education, sports, technology, the environment, the mainstream media, travel, videoblogging and web development and sites. A "default category" is provided for movies which cannot be assigned to neither one of the previous categories. Apart from the video data, the dataset provides associated social metadata, automatic speech recognition (ASR) transcripts and video shot segmentation and key frames. The dataset was successfully validated during 2010-2012 MediaEval benchmarking campaigns [33];

ADL [35]: contains 10 different activities, i.e., answering a phone, dialing a phone, looking up numbers in a phone book, writing on a white board, drinking water, eating a snack, peeling a banana, eating a banana, chopping a

[1] `http://blip.tv/`

banana and eating food with silverware. Each of these activities is performed 3 times by 5 different people. These people have different genders, ethnicity, and appearance so sufficient appearance variation is available in the dataset. Each clip is in the range of 3-50s. In total the dataset contains 150 videos;

VSD2013 [36]: it contains violence annotations for 25 typical Hollywood productions. Movies range from very violent ones (e.g., Saving Private Ryan with 34% violent frames) to movies with (almost) no violence (e.g., Dead Poets Society with less of 1% of violent frames). This dataset (in its various versions) has been exploited during the 2011-2014 MediaEval benchmarking campaigns [36];

UCF101 [34]: consists of 13,320 realistic videos from YouTube[2] with large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. The videos in 101 action categories are grouped into 25 groups, where each group can consist of 4-7 videos of an action. The videos from the same group may share some common features, such as similar background, similar viewpoint, etc. The action categories can be divided into five types: human-object interaction, body-motion only, human-human interaction, playing musical instruments and sports.

These datasets are particularly challenging due to the diversity of video footage and specifically the variability of videos within the same categories. Some examples are illustrated in Figure 4.

4.2 Metrics

To assess performance, depending on the dataset, we employed several standard metrics. We compute the system *accuracy* which represents the number of items correctly classified (true positives + true negatives). To provide a global measure of performance, we estimate the overall *Mean Average Precision* (MAP), which is computed as the mean of the average precision scores for each item:

$$\mathrm{MAP} = \sum_{\mathrm{q}=1}^{\mathrm{Q}} \frac{\mathrm{AP(q)}}{\mathrm{Q}} \tag{7}$$

where Q represents the number of items (queries), and $AP()$ is given by

$$\mathrm{AP} = \frac{1}{\mathrm{m}} \sum_{\mathrm{k}=1}^{\mathrm{n}} \frac{\mathrm{f_c(v_k)}}{\mathrm{k}} \tag{8}$$

where n is the number of items, m is the number of items of category c, and v_k is the k-th item in the ranked list $\{v_1, ..., v_n\}$. Finally, $f_c()$ is a function which returns the number of items of class c in the first k items if v_k is of class c and 0 otherwise (we used the trec_eval scoring tool available at `http://trec.nist.gov/trec_eval/`).

[2] `http://www.youtube.com/`

4.3 Content descriptors

Video information is represented with content descriptors. Currently there is a huge amount of literature in this area and covering all the existing techniques is impossible. For evaluation, we selected some of the most representative approaches known to perform well in many benchmarking scenarios [1–4] as well as which are suitable to our experimentation tasks.

We experimented with the following descriptors:

- **visual descriptors:**
  - *HoG features (81 values)* [37] - exploit local object appearance and shape within an image via the distribution of edge orientations. The image is divided into small connected regions (3x3) and for each region a pixel-wise histogram of edge orientations is computed. In the end, the combination of these histograms represents the final descriptor;
  - *color naming histogram (11 values)* [38] - describes the global color contents and it maps colors to 11 universal color names. We select this feature, instead of the classic color histogram, because the color naming histogram is designed as a perceptually based color naming metric that is more discriminative and compact.
- **motion descriptors:**
  - *HoG-3D (72 values)* [27] - computes HoG features in 3D blocks with a dense sampling strategy: first the gradient magnitude responses in horizontal and vertical directions are computed. Then, for each response the magnitude is quantized in k orientations, where $k = 8$. Finally, these responses are aggregated over blocks of pixels in both spatial and temporal directions and concatenated;
  - *Histograms of optical Flow (72 values)* [40] - computes a rough estimate of velocity at each pixel given two consecutive frames. We use optical flow at each pixel obtained using the Lucas-Kanade method [40] and apply a threshold on the magnitude of the optical flow, to decide if the pixel is moving or is stationary. We divide the frames in 3x3 regions and then we compute the HoF feature for each region [39];
  - *Body-part features (144 values)* [43] approximate the optical flow that is computed on the body-part components. Human pose and body-part motion obtained good results in many event detection categories [41–43]. We extract the body-part components using the state-of-the-art body-part detector [41] and compute at every frame for all 18 body-parts a Histogram of optical Flow in 8 orientations [43];
- **audio descriptors:**
  - *standard audio features (196 values)* [44] - we use a set of general-purpose audio descriptors, namely: Linear Predictive Coefficients, Line Spectral Pairs, MFCCs, Zero-Crossing Rate, spectral centroid, flux, rolloff and kurtosis, augmented with the variance of each feature over a certain window (we use the common setup for capturing enough local

context that is equal to 1.28s). For a sequence, we take the mean and standard deviation over all frames.

Different descriptor combinations were employed depending on the dataset and available information. For *Blip10000* and *VSD2013* we use all the visual and audio features. For these datasets, we decided not to use motion features because of their high computational complexity which makes them inefficient for large size collections. For *UCF101* we only use several of the visual descriptors: HoG to account for feature information, color naming histogram to account for color information and motion features which are representative for this dataset. We did not use audio and text information because the movies from *UCF101* dataset do not contain this information. For the *ADL* dataset we use only the body-part features which already provided state-of-the-art results in many approaches [43].

5 Experimental results

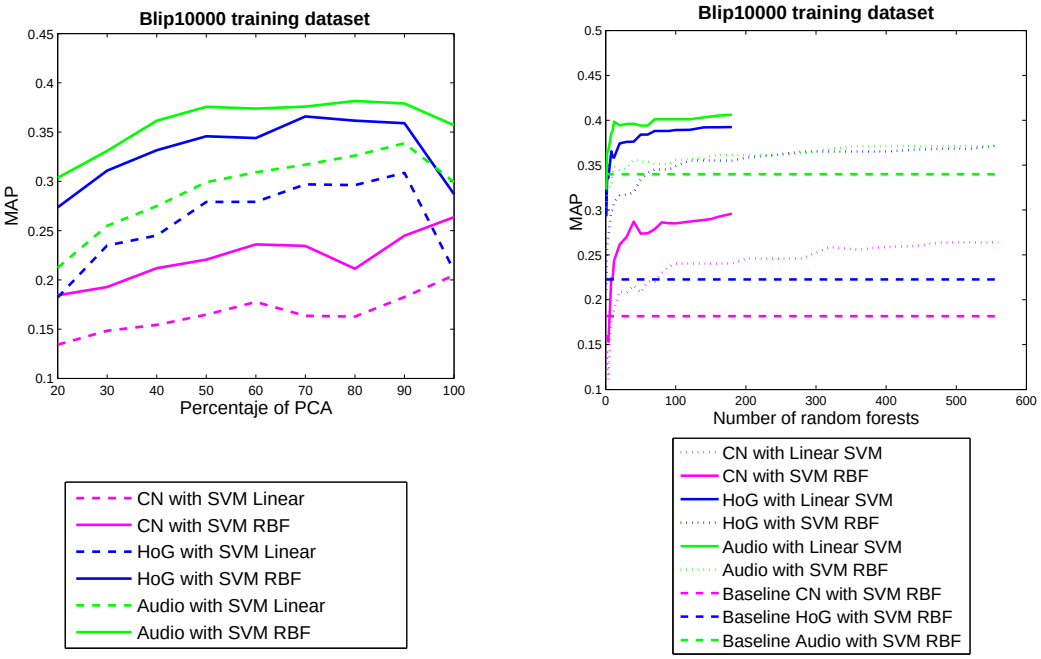
In this section we present and discuss the experimental results for each of the considered classification scenarios.

5.1 Video genre classification

In the first experiment, we test our video content description framework in the context of video genre classification using the *Blip10000* [32] dataset. All the parameter optimization is carried out on the training set which we split in two fixed, equally sized parts, one for training and the other for testing in the context of parameter optimization. We compare our method with the state-of-the-art using the official training set (5,288 videos) and test set (9,550 videos). The performance is measured with MAP. For this experiment, we use the following descriptors (see Section 4.3): visual features (HoG, CN), motion features (HoG-3D) and the standard audio features.

Parameter tuning. In order to refine our parameters, we start with the following baseline setting: 10 random forests which is a good trade-off between speed and accuracy of the results, L2 with Power Normalization for the modified VLAD features and SVMs classifier with linear and RBF kernel.

The first experiment evaluates the influence of Principal Component Analysis (PCA) on system performance. We have two main reasons to make this experiment: firstly, when the number of RF increases the feature length may become very long and we want to make it shorter, and secondly, we expect that PCA will improve the performance by removing the feature noise. Theoretically, the classification approaches work better when the noise is reduced and the data are uncorrelated. Figure 5(a) presents some of the results. One may observe that the PCA improves the performance of HoG features when we keep 80%-90% from the PCA components. However, by applying the PCA on



(a) MAP vs. percentage of PCA reduction

(b) MAP vs. the number of random forests

Fig. 5: Parameter tuning on the training set of *Blip10000* [32] dataset.

CN features, the performance will decrease because the information provided by CN is already uncorrelated. The experiments presented in Figure 5(a) show that we obtain the best performance by reducing the HoG and audio features by 20%. In all the following experiments will use this combination: HoG and audio with PCA and CN without PCA.

In the second experiment we analyze the influence of the number of random forests (see Section 3.2). Some of the results are presented in Figure 5(b). One can observe that the performance increases with increasing the number of random forests. In case of using SVM with nonlinear kernel, the performance plateaus after 100 random trees for all the features. A big improvement can be noticed compared to baseline, i.e., the simple average of the features: CN goes from 0.16 MAP to 0.28 and HoG from 0.29 to 0.38. The proposed approach significantly improves the results. The final sizes of the Fisher vectors are reasonable at 4,400 for CN, 5,600 for HoG and 6,000 for audio. On the other hand, it can be observed that using the linear SVM classifier the performance plateaus after 100 random trees for all the features. However, the RBF kernel obtains an increase of performance of 0.02 over the linear kernel. Therefore, we decide to use only the SVM with nonlinear RBF kernel.

In the third experiment we study the influence of the tree pruning via the t_{stop} parameter (see Section 3.2). The results are presented in Figure 6. For

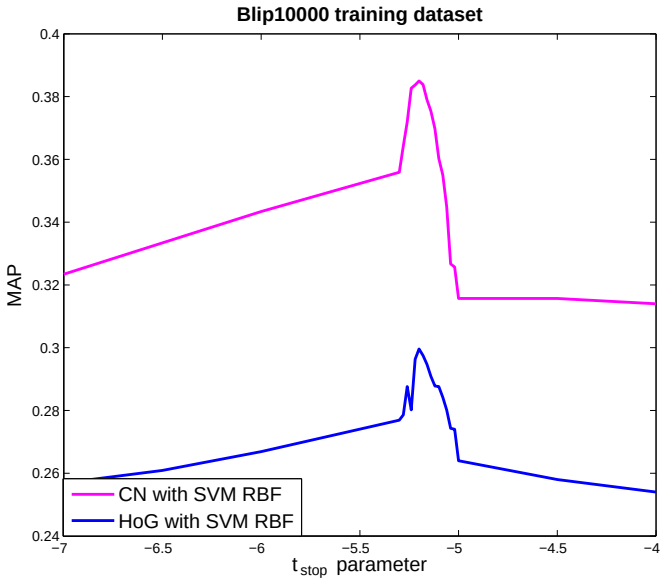


Fig. 6: Parameter tuning on the training set of *Blip10000* [32] dataset: MAP vs. the t_{stop} parameter.

small values of t_{stop}, trees are almost left unpruned and therefore the resulting features' length is very high, leading to a performance drop of more than 5%. On the other hand, when t_{stop} is too high, trees are over pruned (close to the root) leading to short descriptors that are less representative and consequently the performance drops by more than 6 percents.

Comparison with the baseline versions. In this experiment, we compare the performance of our approach with a standard Bag-of-Words, VLAD approach and the simple aggregation of frames (by computing the mean and the dispersion over all the frame). We also present the performance of our approach without Random Forests, when we use the classical k-means for the word assignment step. The goal of this experiment is to demonstrate that the inclusion of both processing steps, i.e., Random Forest based word assignment and the modified VLAD, are crucial for achieving the best performance. Experiments are conducted by adopting the previous parameter tuning on the *Blip10000* testing dataset. The results are presented in Table 1.

The lowest performance (from 0.182 to 0.223) is obtained when we aggregate all the features in one descriptor by computing the mean and the standard deviation over all the frames. However, this is somehow expected as temporal information is lost. Even when mapping multiple vectors into fixed length representations, e.g., using Bag-of-Words or the standard VLAD representation, the performance is still lower with more than 3% for CN and 7% for the

Table 1: Comparison with the baseline versions of the proposed approach - MAP values (*Blip10000* [32] dataset; the best results are presented in bold).

Approach/feature type	*HoG*	*CN*
feature average and SVM RBF	0.182	0.223
Bag-of-Words and SVM RBF	0.232	0.263
VLAD and SVM RBF	0.254	0.314
proposed with k-means and SVM RBF	0.245	0.316
proposed with SVM RBF	**0.295**	**0.385**

VLAD representation. Finally, the proposed algorithm for the word assignment obtains a performance improvement of more than 5% over the standard assignment (with k-means).

Table 2: Comparison with state-of-the-art (*Blip10000* [32] dataset).

approach	*MAP*	*proposed*	*MAP*
block-based audio features and 5-nearest neighbor [45]	0.192	proposed with audio descriptors	0.472
visual color, texture and rgbSIFT descriptors [46]	0.350	proposed with visual descriptors	0.453
-	-	proposed with motion descriptors (HoG-3D features)	0.483
visual and audio descriptors with Fisher kernel [22]	0.55	proposed with visual and audio descriptors	0.533
visual and motion descriptors [48]	0.452	proposed with visual and motion descriptors	0.538
-	-	proposed with audio, visual and motion descriptors	0.571
BoW on text ASR and metadata [47]	0.523	-	-

Comparison to state-of-the-art. In this section we compare our approach against state-of-the-art results from the literature; in particular from MediaEval 2012 benchmarking [33]. For the audio modality, we achieve 0.472 MAP, much better than the best audio only MediaEval result [45] MAP 0.19. For visual modality, at 0.453 MAP we perform significantly better than the best MediaEval result that was obtained exploiting only visual information [46], MAP 0.35. Also, for motion modality (HoG-3D features), we obtained 0.483 MAP which is higher than audio and visual features alone.

Textual data are by far the most representative for providing content information. Specific keywords, e.g., "religion", "economy", "music", can reveal meaningful information about genres. For instance, metadata usually contains the video title, user tags, comments and content descriptions that are highly correlated to genre concepts. Even if the metadata contains very highly semantic information, the main drawback of these features is that they cannot be generated automatically, which limits their applicability. Remarkably, our combination of audio, motion and visual features yield a MAP of 0.571 which

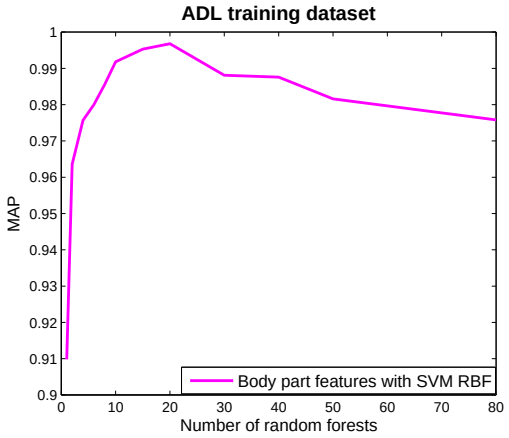


Fig. 7: Parameter tuning on the training set of *ADL* [35] dataset: accuracy vs. the number of random forests.

is higher than the one obtained using highly semantic metadata information, which ranked highest in th MediaEval 2012 benchmarking [47] (MAP 0.523).

We also compare our method with other state-of-the-art approaches, namely: authors in [48] propose a manifold learning based on reciprocal neighborhood and authority of ranked lists for improving retrieval of videos according to their genre. They combine visual features (Bag-of-Visual-Words and Bag-of-Scenes) with motion features (histogram of motion patterns); and authors in [22] who propose the use of Fisher Kernels to model variation in time for frame-based video features. Results are presented in Table 2. One can observe that the proposed approach is able to provide superior performance.

5.2 Daily activities classification

The results of our analysis on the *Blip10000* dataset indicate that the proposed approach obtains good results for the standard problem of video genre classification. A natural question that arises is whether these features also generalize to other datasets and class categories. We examine this question in detail by performing experiments on the *ADL* [35] dataset. Performance is measured in accuracy. We do all the optimizations on the half of the dataset (75 videos) and we report the final results on the full dataset (150 videos).

We did not use HoG and CN features because for this task the contextual information is not relevant. All the human activities happen in an indoor space, having similar background. As human pose and body-part motion are important for distinguishing the different categories, we extract body-part features as presented in Section 4.3.

Parameter tuning. We found no improvements by doing PCA on the body-part HoF features. Figure 7 plots the accuracy with respect to the number of

random forests. Using only a single RF yields an accuracy of 91%. The best accuracy of 99% is obtained using 20 random forests. Note that the number of random forests is relatively low, likely due to the reduced number of videos in the dataset. At 20 random forests the final feature has 11,520 dimensions. These settings are adopted for the following experiment.

Comparison with the baseline versions. In this experiment, we compare the performance of our approach with a standard Bag-of-Words, VLAD approach, simple aggregation of frames (by computing the mean and the dispersion over all the frame) and our approach with a classical k-means for the word assignment step instead of the Random Forests. Experiments are conducted by adopting the previous parameter tuning on the ADL testing dataset. The results are presented in Table 3.

The lowest performance (from 89.2% to 93.3%) is obtained when we aggregate all the features in one descriptor and with the Bag-of-Words representation. The performance of the standard VLAD representation is still lower with more than 1%, while the proposed algorithm for the word assignment obtains a performance improvement of more than 2% over the standard assignment with k-means.

Comparison to State-of-the-Art. We compare the proposed approach with others from the literature (see Table 4). As it can be observed, our approach yields the highest accuracy of 99.3%. The results clearly show that our representation enhances the discriminative power of features and improves the action recognition performance. We conclude that the proposed representation is also effective for modeling the frame based body-part features.

Table 3: Comparison with the baseline versions of the proposed approach - accuracy values (*ADL* [35] dataset; the best results are presented in bold).

Method	*Accuracy*
feature average and SVM RBF	89.2%
Bag-of-Words and SVM RBF	93.3%
VLAD and SVM RBF	98.1%
proposed with k-means and SVM RBF	97.4%
proposed with SVM RBF	**99.3%**

Table 4: Comparison with state-of-the-art (*ADL* [35] dataset; the best results are presented in bold).

Approach	*Description*	*Accuracy*
Bilinski et al. [49]	BoW of relative dense tracklets	92.0%
Raptis et al. [51]	BoW of spatio-temporal tracklets	94.5%
Wang et al. [50]	Multiscale Spatio-Temporal Contexts with Multiple Kernel Learning	96.0%
Rostamzadeh et al. [43]	Body-part HoF with Fisher kernel	98.75%
proposed	Body-part HoF with modified VLAD	**99.3%**

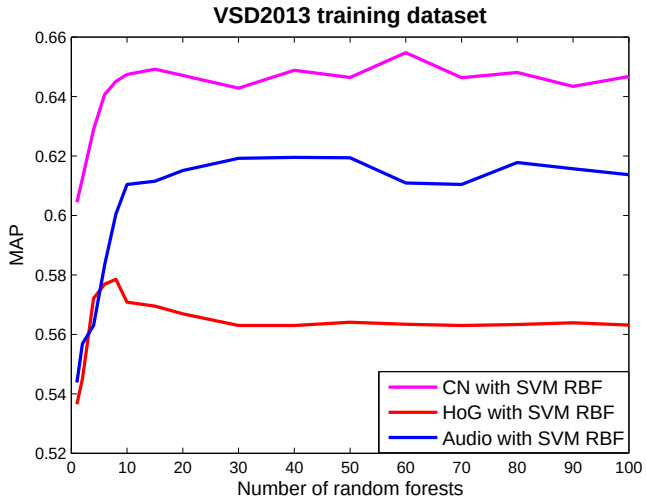


Fig. 8: Parameter tuning on the training set of *VSD2013* [36] dataset: MAP vs. the number of random forests.

5.3 Violent scenes classification

We also test our approach on the 2013 Violence Scenes Detection *VSD2013* dataset [36]. This data is annotated at video shot level for the presence of violence (yes/no annotations). Two different use cases are annotated: an objective definition of violence where violence is defined as "physical violence or accident resulting in human injury or pain" and a subjective one where the targeted violent segments are those "one would not let an 8 years old child see in a movie because they contain physical violence". Both scenarios were tested in our experiments.

We do all the optimizations on the official training dataset that contains 18 of the movies (a total of 32,678 video shots), while the actual benchmarking is carried out on the remaining 7 movies (11,245 shots). Performance is assessed using MAP.

For this task we use the following features: HoG, CN and audio features (see Section 4.3).

Parameter tuning. We first optimize the dimension reduction using PCA. We found that the performance of audio features is not improved with PCA. Only for HoG and CN we obtain a good improvement by reducing the dimension to 90%.

Figure 9 plots the accuracy with respect to the number of random forests. It can be observed that the results are increasing while using a higher number of random trees. The performance plateaus however after 12 random forests. The best accuracy is obtained with the CN descriptor and audio features (intuitively, violence is highly correlated to the color and audio information).

Table 5: Comparison with the baseline versions of the proposed approach - MAP values (*VSD2013* [36] dataset - objective task; the best results are presented in bold).

Method	*MAP*
feature average and SVM RBF	0.6131
Bag-of-Words and SVM RBF	0.6634
VLAD and SVM RBF	0.6915
proposed with k-means and SVM RBF	0.7011
proposed with SVM RBF	**0.7202**

Table 6: Comparison with state-of-the-art (*VSD2013* [36] shot based classification; the best results are presented in bold).

Method	*Description*	*Accuracy*
Objective annotation		
FAR [54]	Multi Layer Perceptron with aural-visual frame features	0.3504
TUDCL [53]	Multiple Kernel Learning with temporal, audio and visual features	0.4202
proposed	HoG frame features with modified VLAD	0.5601
proposed	Audio frame features with modified VLAD	0.6137
proposed	CN frame features with modified VLAD	0.6695
proposed	Audio-visual frame features with modified VLAD	**0.7202**
Subjective annotation		
TECH-INRIA [52]	Bayesian networks with temporal, audio and visual features	0.4479
proposed	HoG frame features with modified VLAD	0.7206
proposed	Audio frame features with modified VLAD	0.6276
proposed	CN frame features with modified VLAD	0.7206
proposed	Global frame features with modified VLAD	**0.7612**

On the other hand, the lowest performance is obtained with the HoG features, which may be due to the fact that violence is not correlated with the type of objects that are part of the scene. We set therefore to 12 random forests.

Comparison with the baseline versions. In this experiment, we compare the performance of our approach with other baseline approaches. Experiments are conducted by adopting the previous parameter tuning on the VSD2013 testing dataset. The results are presented in Table 5.

The performance values are similar with those obtained in the previous experiments. The simple feature aggregation and the Bag-of-Words representation leads to MAP values from 0.6131 to 0.6634. Also, the performance of standard VLAD representation is still lower with almost 3%. Finally, the proposed algorithm for the word assignment obtains a performance improvement of more than 2% over the standard assignment with k-means.

Comparison to State-of-the-Art. We compare our approach against the results obtained at the MediaEval 2013 benchmarking [55]. Given the classification nature of our approach, we compare in particular to the shot-based violence classification results [56].

Table 7: Comparison with the baseline versions of the proposed approach - accuracy values (*UCF101* [34] dataset; the best results are presented in bold).

Method	*Accuracy*
feature average and SVM RBF	67.2%
Bag-of-Words and SVM RBF	68.1%
VLAD and SVM RBF	73.1%
proposed with k-means and SVM RBF	73.6%
proposed with SVM RBF	**74.1%**

A summary of the best team runs is presented in Table 6 (results are presented by decreasing MAP values). The most efficient approaches remain those that include multimodal information (e.g., motion, visual and aural) and aggregate it with several late fusion techniques. The best MAP value is 0.42 for the objective annotations [53], while a similar value of 0.447 is obtained also on the subjective annotations [52]. For using visual features only, at 0.6695 MAP (objective) and 0.7206 (subjective) we perform significantly better than the best results. Remarkably, our combination of audio and visual features yields with more than 0.04 MAP better than the use of individual modalities. We conclude that the proposed approach improves the retrieval performance for all modalities, outperforming other state-of-the-art approaches.

5.4 Human action classification

Finally, we also test our approach on a human action classification task. For this purpose, we report results on the *UCF101* [34] dataset. Performance is evaluated in terms of classification accuracy. We perform all optimization on a quarter of the dataset (3,207 videos). We then compare with the state-of-the-art using the full dataset (13,320 clips) [34].

We use the following features: HoG, HoF and CN (with 3x3 spatial division; see Section 4.3).

Parameter tuning. We first optimized the dimension reduction using PCA. We found that both the CN descriptor and the HoF did not benefit from applying PCA reduction, only for HoG we obtain a good improvement by reducing dimensions to 90%.

In Figure 9 we evaluate the performance with respect to the number of random forests. For HoF we obtain the best results when we use more than 80 random trees, while for CN and HoG we obtain the best results when we use only a small number of random forests. In the next experiment, we set 10 random trees for CN and HoG, and 150 for the HoF features.

Comparison with the baseline versions. In this experiment, we compare the performance of our approach with other baseline approaches. Experiments are conducted by adopting the previous parameter tuning on the UCF101 testing dataset. The results are presented in Table 7.

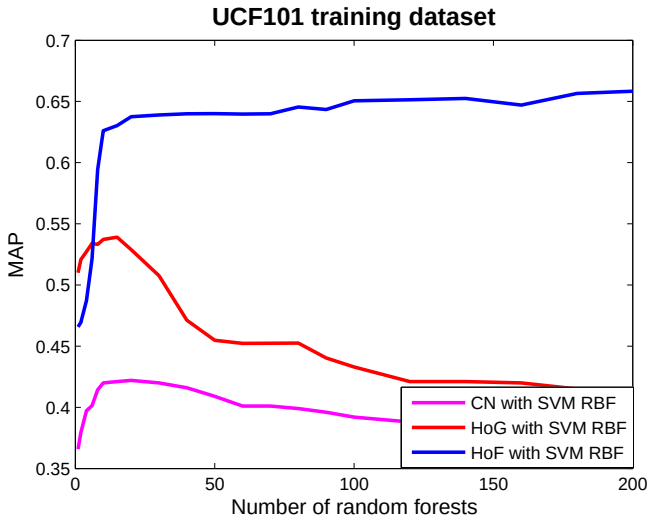


Fig. 9: Parameter tuning on the training set of *UCF101* [34] dataset: accuracy vs. the number of random forests.

The obtained performance is similar with the results obtained in the previous baseline experiments. The proposed approach obtains better performance than the other baseline approaches.

Comparison to State-of-the-Art.

In this section we compare our method with other state-of-the-art approaches: in [34] the authors use a set of standard local motion features, such as STIPs and dense trajectory features; in [59] authors propose a new technique to match dense trajectories and remove those that contain background motion noise; authors in [57] present an approach that uses Convolutional Neural Networks; in [58] the authors proposed a method that combines the dense trajectories with a classical VLAD encoding; authors in [60] apply a two-stream ConvNet architecture which incorporates spatial and temporal networks with multi-frame dense optical flow. We also compare our method with the results achieved at the THUMOS challenge [61] which employed the UCF101 datasets. Results are presented in Table 8.

As it can be seen, the proposed approach achieved an accuracy of 74.1%. The highest accuracy is obtained by Simonyan et al. [60], 87.90%, with discriminatively trained deep Convolutional Networks but at the price of a significantly higher computational complexity. We still obtain better results than Soomro et al. [34], Jain et al. [58], Karpathy et al. [57] and Murthy et al. [59], which obtain 43.90%, 52.10%, 65.40% and 73.1% respectively. However, by using the frame-based features, we obtain lower results than the best results of the THUMOS competition: [62] - authors propose a framework that incorporates the dense trajectories (HoG-3D / HoF-3D / MBH), spatio-temporal pyramids with a modified version of Fisher Kernel representation; [63] - authors

Table 8: Comparison with state-of-the-art (*UCF101* [34] dataset; the best results are presented in bold).

Method	*Description*	*Accuracy*
Soomro, et al. (2012) [34]	Cuboid descriptors	43.90%
Jain, et al. (2013) [58]	dense trajectories + VLAD encoding	52.10%
Karpathy, et al. (2014) [57]	Convolutional Neural Networks	65.40%
Murthy, et al. (2013) [59]	Ordered trajectories + VLAD encoding	73.10%
proposed approach	Global frame features with modified VLAD	74.10%
Karaman, et al. (2013) [63]	P-SIFT, P-OSIFT, HoG-3D / HoF-3D / MBH with BOW	85.70
Wang, et al. (2013) [62]	dense trajectories (HoG-3D / HoF-3D / MBH) with a modified FK	85.90%
Simonyan, et al. (2014) [60]	ConvNet architecture	**87.90**%

propose a Bag-of-Features pipeline in combination with local SIFT pyramids (P-SIFT), opponent color keyframes (P-OSIFT), HoG-3D / HoF-3D / MBH. These methods outperform the proposed method with more than 10%. However, the advantage of our approach is in the use of simple global features, whereas all the other better performing methods use computationally more expensive Space-Time Interest Points (trajectories) or more complex architectures, such as deep learning. Also, we conclude that the frame-based features are not the best approach for action recognition tasks, and the use of local dense trajectories may lead to better performance. However, our framework yields good performance while using simpler features.

5.5 Computational complexity

In this section we discuss the computational complexity of the proposed description framework. We analyze the time for computing each processing step, from feature extraction to video classification. We perform this experiment on the *Blip10000* [32] dataset which contains more than 1TB of video information and up to 2,000 hours of video footage. The run-time is evaluated on a regular PC machine using a 2.9 GHz Intel Xeon CPU and 24GB of RAM. We do not use parallelization. Experiments were run with HoG and CN features, with 100 random forests and L2 with square-root normalization.

The computational cost per frame is presented in Figure 10. Descriptor extraction takes 200 milliseconds (ms) per image (180 ms for HoG and 20 ms for CN). The input/output operation lasts more than 11% of the global computation time (30 ms per frame). The VLAD computation is very fast, namely 12 ms per frame. Finally, classification takes 17 ms for all classes. Therefore, a processing chain for HoG would take 239 ms per frame ($\sim$ 6 seconds for 1 second of video, i.e., 25 frames) while for CN 79 ms per frame ($\sim$ 2 seconds for 1 second of video).

We conclude that this represents a reasonable, near real-time, cost considering the achieved performance. This is achieved without any algorithm optimization nor adequate hardware acceleration or parallel implementation.

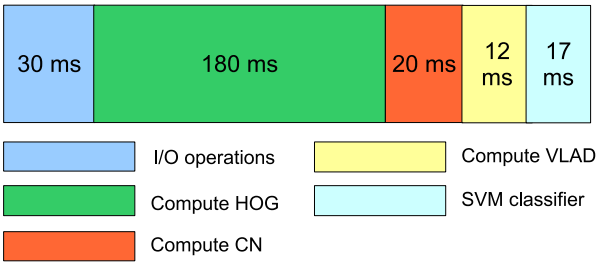


Fig. 10: Total computational time (ms) per frame for the proposed video description framework (*Blip10000* [32] dataset, use of HoG and CN features).

Using clustering processing will allow to easily achieve faster than real-time performance.

Finally, we also compare our implementation (modified VLAD with frame based features) against other approaches from the literature. Results are presented in Table 9.

First of all, we show that frame-based features are more computational efficient than classical spatio-temporal features. Therefore, we selected as example a system that integrates the proposed modified VLAD approach with spatio-temporal HoG-3D features [27]. In this case, spatio-temporal feature extraction required more than 800ms. In contrast, frame-based feature extraction takes only 200ms. In addition, for each frame the number of generated spatio-temporal features is greater than for the frame-based approach, which impacts the performance of the VLAD algorithm. In effect, the resulting VLAD approach can become even ten times slower.

We also compare the efficiency of the proposed modified VLAD approach against other approaches that use similar frame-based features, e.g., the approach in [22] which uses Fisher Kernels to model variation in time for frame-based video features and a standard VLAD implementation [28]. We achieve a total processing time of 219ms and 221ms, respectively. These results show that by performing the fast word assignment with the pruned random forest trees, the proposed algorithm is capable of achieving lower processing time, i.e., 212 ms.

6 Conclusion

We proposed a new video representation framework that models the variation in time. It uses a fast word assignment approach by replacing Bag-of-Words k-means visual vocabulary assignment with a Random Forest approach. A modified version of Vector of Locally Aggregated Descriptor with Fisher Kernel representation is then used for increasing the representative power of the descriptors.

Table 9: Comparison of computational efficiency (experiments were conducted on the *Blip10000* [32] dataset).

Approach	*Feature extraction*	*Feature aggregation*	*Total time*
proposed approach with spatio-temporal HoG-3D [27] features	833ms	150ms	988ms
Frame-based features + Fisher kernel + SVM [22]	200ms	19ms	219ms
Frame-based features + classical VLAD [28] + SVM	200ms	21ms	221ms
proposed approach with frame-based features + SVM	200ms	12ms	212ms

We demonstrated that our framework is highly general: we showed significant improvements on a wide variety of features, ranging from global visual features, body-part features, audio. In order to combine all these modalities, we used a late fusion strategy that significantly improved the performance of the system, which makes the system to be easier for scaling up. We also showed that our method works on a wide variety of classification scenarios: we obtained good performance on action classification (UCF101 dataset) using global features instead of the more complex STIPs or dense trajectories used in other methods. We also improved the state-of-the-art on daily activities classification (ADL dataset) and we significantly improved the state-of-the-art on video genre classification (Blip10000 dataset) and violent scenes classification (VSD2013 dataset). On Blip10000 dataset, we prove that notwithstanding the superiority of employing user-generated textual information (e.g., user tags, metadata), the proposed multimodal approach obtained higher performance than the textual descriptors.

Future work will mainly address the extension of this approach to integrating with low computational complexity the more demanding to compute spatio-temporal information. Also, an interesting lead would be to study the impact of more elaborated late fusion approaches in the current framework.

Acknowledgements The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132395.

References

1. P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A.F. Smeaton, G. Quéenot, TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics, Proceedings of TRECVID 2013, `http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/tv13overview.pdf`, NIST, USA, 2013.
2. A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, H. Müller, Overview of the ImageCLEF 2013 medical tasks, Working Notes of CLEF 2013 (Cross Language Evaluation Forum), Valencia, Spain, 2013.

3. MediaEval 2013 Workshop, Eds. M. Larson, X. Anguera, T. Reuter, G.J.F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, M. Soleymani, co-located with ACM Multimedia, Barcelona, Spain, October 18-19, CEUR-WS.org, ISSN 1613-0073, Vol. 1043, http://ceur-ws.org/Vol-1043/, 2013.
4. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.
5. D. Brezeale, D. J. Cook, "Automatic video classification: A survey of the literature", in Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 38(3), 416-430, 2008.
6. X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. CoRR , abs/1405.4506, 2014
7. H. Wang, A. Klaser, C. Schmid, and C. Liu. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, vol. 103(1), pp. 60-79, 2013.
8. H. Wang and C. Schmid. Action recognition with improved trajectories. In Proc. ICCV, pp. 3551-3558, 2013.
9. Z. Ma, Y. Yang, N. Sebe, and A. Hauptmann, Knowledge Adaptation with Partially Shared Features for Event Detection Using Few Exemplars, IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(9):1789-1802, September 2014
10. K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Fisher networks for large-scale image classification. In NIPS , 2013.
11. G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, Visual Categorization with Bags of Keypoints, European Conference on Computer Vision (ECCV), pp. 1-2, 2004.
12. J.R.R. Uijlings, A.W.M. Smeulders, R.J.H. Scha, Real-Time Visual Concept Classification, In IEEE Transactions on Multimedia, vol. 12(7), pp. 665-681. 2010.
13. F. Perronnin, J. Sanchez, T. Mensink, "Improving the fisher kernel for large-scale image classification", in European Conference of Computer Vision (ECCV), pp. 143-156, 2010.
14. N. Ikizler-Cinbis, S. Sclaroff, "Object, scene and actions: combining multiple features for human action recognition", in Proceedings of the European Conference on Computer vision (ECCV), pp. 494-507, 2011.
15. J. Liu, J. Luo, M. Shah, "Recognizing realistic actions from videos in the wild", in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1996-2003, 2009.
16. A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in Advances in Neural Information Processing Systems (NIPS), pp. 1097-1105, 2012.
17. D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification", In Proceedings-International Joint Conference on Artificial Intelligence (IJCAI), vol. 22(1), pp. 1238-1242, 2011.
18. V. Quoc, "Building high-level features using large scale unsupervised learning", in IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP), 2013.

19. I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008), pp. 1-8, 2008.
20. B. Chakraborty, M. B. Holte, T. B. Moeslund, J. Gonzlez, Selective spatio-temporal interest points, in Computer Vision and Image Understanding, 116(3), 396-410, 2012.
21. B. Solmaz, S. M. Assari, and S. Mubarak, "Classifying web videos using a global video descriptor". Machine vision and applications (MVAP), vol. 24(7), pp. 1473-1485, 2013.
22. I. Mironica, J. Uijlings, N. Rostamzadeh, B. Ionescu, N Sebe, "Time matters!: capturing variation in time in video using fisher kernels", in Proceedings of the 21st ACM International Conference on Multimedia, pp. 701-704, 2013.
23. L. Breiman, "Random forests", in Machine Learning, vol. 45(1), pp. 5-32, 2001.
24. A. Bosch, A. Zisserman, X. Munoz, "Image classification using random forests and ferns", in IEEE International Conference on Computer Vision (ICCV), 2007.
25. J. Marin, D. Vzquez, A. M. Lpez, J. Amores, B. Leibe, "Random Forests of Local Experts for Pedestrian Detection", in IEEE International Conference on Computer Vision (ICCV), pp. 2592-2599, 2013.
26. A. Zimek Arthur, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data.", in Statistical Analysis and Data Mining, vol. 5.5, pp. 363-387, 2012.
27. J.R.R. Uijlings, I.C. Duta, E. Sangineto, N. Sebe, "Video classification with Densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off", in International Journal of Multimedia Information Retrieval, pp. 1-12, 2014.
28. H. Jégou, M. Douze, C. Schmid, P. Pérez. "Aggregating local descriptors into a compact image representation", in Computer Vision and Pattern Recognition (CVPR), 2010.
29. H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2012.
30. C. Imre, J. Korner, "Information theory: coding theorems for discrete memoryless systems", Cambridge University Press, 2011.
31. R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, D. Scuse, WEKA Manual for Version 3-7-8, 2013.
32. S. Schmiedeke, P. Xu, I. Ferrané, M. Eskevich, C. Kofler, M. Larson, Y. Estève, L. Lamel, G. Jones, T. Sikora, "Blip10000: A Social Video Dataset Containing SPUG Content for Tagging and Retrieval", ACM Multimedia Systems Conference, February 27 - March 1, Oslo, Norway, 2013.
33. S. Schmiedeke, C. Kofler, I. Ferrané, "Overview of the MediaEval 2012 Tagging Task", Working Notes Proceedings of the MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073, `http://ceur-ws.org/Vol-927/mediaeval2012_submission_2.pdf`.
34. S. Khurram, A. R. Zamir, and M. Shah: "Ucf101: A dataset of 101 human actions classes from videos in the wild." CoRR, abs/1212.0402, 2012.
35. R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In ICCV, 2009.

36. C.-H. Demarty, C. Penet, M. Soleymani, G. Gravier, "VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation", Media Tools and Applications, 2013
37. O. Ludwig, D. Delgado, V. Goncalves, U. Nunes: "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection", IEEE Int. Conference On Intelligent Transportation Systems, 1, pp. 432-437, 2009.
38. J. Van de Weijer, C. Schmid, J. Verbeek, D. Larlus: "Learning color names for real-world applications", IEEE Trans. on Image Processing, 18(7), pp. 1512-1523, 2009.
39. S. Lazebnik, C. Schmid, J. Ponce: "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." Computer Vision and Pattern Recognition, 2006.
40. B. Lucas, T. Kanade, "An iterative image registration technique with an application to stereo vision", in Proceedings of Imaging Understanding Workshop, 1981.
41. M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, A database for fine grained activity detection of cooking activities. In International Conference of Computer Vision and Pattern Recognition, CVPR, 2012.
42. Y. Yang, D. Ramanan, "Articulated human detection with flexible mixtures of parts", in Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(12), 2878-2890, 2013.
43. N. Rostamzadeh, G. Zen, I. Mironic, J. Uijlings, N. Sebe, "Daily Living Activities Recognition via Efficient High and Low Level Cues Combination and Fisher Kernel Representation", IEEE International Conference on Image Analysis and Processing, ICIAP, 2013.
44. B.Mathieu, S.Essid, T.Fillon, J.Prado, G.Richard: "YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software", Proceedings of the 11th ISMIR conference, pp. 441-446, 2010.
45. B. Ionescu, I. Mironică, K. Seyerlehner, P. Knees, J. Schluter, M. Schedl, H. Cucu, A. Buzo, and P. Lambert. ARF @ mediaeval 2012: Multimodal video classification. In MediaEval workshop, 2012.
46. T. Semela, M. Tapaswi, H. Ekenel, and R. Stiefelhagen. Kit at mediaeval 2012 - content-based genre classification with visual cues. In MediaEval workshop, 2012.
47. S. Schmiedeke, P. Kelm, and T. Sikora. TUB @ MediaEval 2012 tagging task: Feature selection methods for bag-of- (visual)-words approaches. In MediaEval Workshop, 2012.
48. J. Almeida, D.C. Pedronette, O. A. Penatti. Unsupervised Manifold Learning for Video Genre Retrieval. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer International Publishing, pp. 604-612, 2014.
49. P. Bilinski, E. Corvee, S. Bak, and F. Bremond, "Relative dense tracklets for human action recognition". IEEE International Conference of Automatic Face and Gesture Recognition (FG), 2013.
50. J. Wang, Z. Chen, and Y. Wu. Action recognition with multiscale spatio-temporal contexts. In CVPR, 2011
51. M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis", in European Conference of Computer Vision (ECCV), pp. 577-590, 2011.

52. C. Penet, C.-H. Demarty, G. Gravier, P. Gros, "Technicolor/INRIA Team at the MediaEval 2013 Violent Scenes Detection Task", Working Notes Proc. [3], 2013.
53. S. Goto, T. Aoki, "TUDCL at MediaEval 2013 Violent Scenes Detection: Training with Multimodal Features by MKL", Working Notes Proc. [3], 2013.
54. M. Sjöberg, J. Schlüter, B. Ionescu, M. Schedl, "FAR at MediaEval 2013 Violent Scenes Detection: Concept-based Violent Scenes Detection in Movies", In MediaEval 2014 Workshop, Barcelona, 2013.
55. C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V.L. Quang, Y.-G. Jiang, "The MediaEval 2013 Affect Task: Violent Scenes Detection", Working Notes Proc. [3], 2013.
56. C.-H. Demarty, B. Ionescu, Y.-G. Jiang, V.L. Quang, M. Schedl, C. Penet, "Banchmarking Violent Scenes Detection in Movies", IEEE International Workshop on Content-Based Multimedia Indexing - CBMI 2014.
57. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fe, "Large-scale video classification with convolutional neural networks", In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
58. M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
59. O. V. Murthy and R. Goecke., "Ordered Trajectories for Large Scale Human Action Recognition", IEEE International Conference on Computer Vision, 2013.
60. K. Simonyan, A. Zisserman, "Two-stream convolutional networks for action recognition in videos", in Conference of Computer Vision and Patern Recognition, 2014.
61. Y.-G. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes", ICCV Workshop on Action Recognition with a Large Number of Classes, http://crcv.ucf.edu/ICCV13-Action-Workshop, 2013.
62. H. Wang, C. Schmid. "LEAR-INRIA submission for the THUMOS workshop." ICCV Workshop on Action Recognition with a Large Number of Classes, 2013.
63. S. Karaman, L. Seidenari, A. D. Bagdanov, A. Bagdanov, "L1-regularized logistic regression stacking and transductive CRF smoothing for action recognition in video." ICCV workshop on action recognition with a large number of classes, 2013.
64. S. Nowozin, "Improved information gain estimates for decision tree induction", arXiv preprint arXiv:1206.4620., 2012.
65. K. Gold, A. Petrosino, "Using information gain to build meaningful decision forests for multilabel classification", In Development and Learning (ICDL), 2010 IEEE 9th International Conference on (pp. 58-63). IEEE, 2010.
66. D. Picard, P-H. Gosselin. "Improving image similarity with vectors of locally aggregated tensors" Image Processing (ICIP), 2011 18th IEEE International Conference on. IEEE, 2011.
67. H. Nakayama, "Aggregating Descriptors with Local Gaussian Metrics", in proceedings of NIPS 2012 Workshop on Large Scale Visual Recognition and Retrieval, 2012.