BEYOND BAG-OF-WORDS: FAST VIDEO CLASSIFICATION WITH FISHER KERNEL VECTOR OF LOCALLY AGGREGATED DESCRIPTORS

Ionuţ Mironică[1], *Ionuţ Duţă*[2,1], *Bogdan Ionescu*[1], *Nicu Sebe*[2]

[1]LAPI, University Politehnica of Bucharest, Romania
[2]DISI, University of Trento, Italy
{*imironica,bionescu*}*@imag.pub.ro,* {*duta,sebe*}*@disi.unitn.it*

ABSTRACT

In this paper we introduce a new video description framework that replaces traditional Bag-of-Words with a combination of Fisher Kernels (FK) and Vector of Locally Aggregated Descriptors (VLAD). The main contributions are: (i) a fast algorithm to densely extract global frame features, easier and faster to compute than spatio-temporal local features; (ii) replacing the traditional k-means based vocabulary with a Random Forest approach that allows significant speedup; (iii) use of a modified VLAD and FK representation to replace the classic Bag-of-Words and obtaining better performance. We show that our framework is highly general and is not dependent on a particular type of descriptor. It achieves state-of-the-art results in several classification scenarios.

Index Terms— Fisher Kernel Vector of Locally Aggregated Descriptor, Random Forests, video classification.

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. Video footage is now the largest broadband traffic category on the Internet, e.g., video broadcasting comprises more than a quarter of the total Internet traffic (source CISCO[1]), over 100 hours of video footage are uploaded to YouTube[2] every minute whereas more than 6 billion hours of video are watched monthly on the same platform — that's almost one hour for every person on Earth. This makes video in particular one of the most challenging data to process.

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, 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 [2–4]. However, the main drawback of these methods is in their limited generalization. Most of them are designed to solve a single application and cannot be generalized to a broad category of video classification tasks. Another weak point of current video description approaches is in the limited exploitation of the defining information of video, i.e., temporal information. Local motion features, dense trajectories or spatio-temporal volumes [2] attempt to solve this problem, but at a very high computational cost. This prevents these features for being reliable candidates for real-time categorization scenarios.

In this context, we introduce 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 Vector of Locally Aggregated Descriptors 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 overview the current state-of-the-art and situate our work. Section 3 details the proposed approach. The experimental results are presented in Section 4 while Section 5 concludes the paper and discusses future perspectives.

2. RELATED WORK

The success of video classification relies mainly on the efficiency of the content description scheme. A large category of video descriptors use global descriptions via the aggregation of local spatio-temporal features, e.g., Histograms of oriented Gradients (HoG), Histograms of optical Flow (HoF), Motion Boundary Histograms (MBH) [2]. These are typically encoded into a Bag of Words (BoW) representation. Spatio-

[1]http://www.cisco.com/
[2]http://www.youtube.com/

temporal interest point-based methods [3] represent the scene and the performed actions as a combination of local descriptors, which are computed in a neighborhood of some interest point information. The neighborhood can be selected as an image patch or as a spatio-temporal volume, e.g., cuboid.

Instead of computing local video features over spatio-temporal cuboids, shallow video representations [4] make use of dense point trajectories. They are also typically relying on the 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.

Another perspective is the use of other encoding approaches such as the Fisher vector encoding [5] or Vector of Locally Aggregated Descriptors (VLAD) [9]. For instance, authors in [8] propose the use of VLAD and Fisher kernels. They used several speedup approaches for densely sampling 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, Random Forests based vocabulary or the Fisher encoding.

Other approaches are more focused and exploit human tracking algorithms to describe video scene information [18], e.g., use of human, objects and scene-based features; use of deformable part models; use of HoF features extracted from human body-part segments. These frameworks are however adapted to specific tasks and require foreground estimation and the detection and tracking of the humans in the scene. The performance of the description scheme is actually highly correlated to the performance of the human detectors.

Although not a description approach themselves, it is worth mentioning the Convolutional Neural Networks (CNNs) as models for describing video content. CNNs prove to be highly efficient in video classification task [6] thanks to the capacity of scaling up to tens of millions of parameters and massive labeled data. However, from a computational perspective, CNNs require important training time to effectively optimize the parameters that compose the model. This difficulty is further compounded when addressing in particular video representation.

In this paper we introduce a new video description framework designed specifically for fast video classification tasks. To reduce computational complexity, the proposed approach replaces the traditional k-means visual vocabulary creation with a Random Forest approach and uses a modified Vector of Locally Aggregated Descriptor (VLAD) and Fisher Kernel representation to replace the standard BoW approach. Designed this way, it allows to capture frame-based video content variation in time.

In the context of the current state-of-the-art, we identify the following contributions of this work: (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) we propose a new algorithm for fast word assignment that uses a set of pruned Random Forest trees; (iii) the proposed framework is feature independent in the sense that is not adapted to the use of a particular type of description scheme. It can work basically with any content representation approach, from basic histograms to more complex approaches that include feature points and multimodal integration; (iv) we demonstrate the generality of the proposed framework in terms of applicability (i.e., it is successfully tested on three different classification scenarios); (v) we achieve similar or better performance compared to the state-of-the-art by using simpler and faster to compute descriptors.

3. PROPOSED APPROACH

The proposed description framework consists of four stages. Firstly, frame-based descriptors are computed for the input video data. Then, the following processing is carried out: (1) we create a dictionary of frame words by replacing the traditional k-means used with Bag-of-Words with a Random Forest based approach. This allows for a significant speedup; (2) each video frame is then assigned to the nearest word according to the previously trained random trees. For each individual modality, we compute a Fisher Kernel (FK) representation of Vector of Locally Aggregated Descriptors (VLAD); (3) finally, the actual classification task is performed by a Support Vector Machine (SVM) classifier. This last step acts also as a late fusion mechanism if multimodal descriptors are used.

3.1. Word assignment with Random Forests

Random Forests (RF) [7] are ensemble learning methods that operate by constructing a multitude of decision trees in the training stage and combine their classification outputs. Given their architecture, RF are in particular good candidates for speeding up the word assignment phase, in particular: (i) have good classification accuracy proved in many scenarios; (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; (iii) RF trees are independent of the descriptor dimension as only one dimension is selected for choosing the path in the tree.

Inspired by [8], we propose the following improved scheme. We use a set of Random Forests that represent a combination of decision trees, where each tree is built independently on the same input descriptors, S. In this configuration, tree leaves correspond to the clusters. Trees are constructed in an recursive way. For each node, a number of split offs are proposed by selecting at random one dimension of the descriptors and a random threshold. Each node splits the descriptors in two sets: S_L (left) and S_R (right). The process is repeated until a leaf is reached (cluster). This corresponds to

the *training phase*. In the *classification part*, descriptors are assigned to each (pre-trained) tree. This conducts to a certain path and a corresponding cluster. Then, the video descriptor is recomputed with the proposed FK and VLAD representation (see Section 3.2), achieving a content representation per tree. The final descriptor is the simple combination of each individual tree representations.

Following this representation scheme, the estimated dimension of the resulting descriptor is given by: $2^{depth} \cdot \#Trees \cdot n \cdot 2$, where $depth$ is the depth of the trees, $\#Trees$ is the number of trees for the Random Forests and n is the initial descriptor dimension (the final multiplication by 2 comes from the FK and VLAD representation, see Section 3.2). In general, results show that a high classification accuracy is achieved with a large number of trees and a higher depth within the trees. However, in this case, descriptors become significantly large, e.g., for only 10 trees of depth 10 and 72 dimensional input descriptors it achieves 1,474,560 values.

We propose a novel way to decrease directly the dimensionality without any additional processing steps and while maintaining the same accuracy. The idea reposes on the definition of the information gain for a node, I, using Shannon's entropy:

$$I = -\frac{\#(S_L)}{\#(S)} H(S_L) - \frac{\#(S_R)}{\#(S)} H(S_R) \tag{1}$$

where $\#(.)$ denotes the cardinality of a set, S is the initial set of descriptors, and H is the Shannon's entropy of the class labels of the descriptors, given by [11]:

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i) \tag{2}$$

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.

Having this information, we now stop the tree from growing by removing the branches that have an information gain lower than a certain threshold. 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 (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.

3.2. Fisher Kernel and VLAD representation

Results from the literature show that adopting Fisher Kernel [8] and VLAD [10] representations allow for achieving higher accuracy than the use of traditional Bag-of-Words histogram representations.

In this context, we propose a new way of representing the descriptor information that exploits the advantages of both, FK and VLAD representations, in a unified framework. The proposed descriptor is represented as the concatenation of the $v_{\mu,i}$ and $v_{\sigma,i}$ FK representations, for $i = 1, ..., K$, with K the number of words (i.e., clusters):

$$v_{\mu,i} = \frac{1}{T\sqrt{P(x_i)}} \sum_{t=1}^{T} \frac{(x_t - \mu_i)}{\sigma_i} \tag{3}$$

$$v_{\sigma,i} = \frac{1}{T\sqrt{2P(x_i)}} \sum_{t=1}^{T} \left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right] \tag{4}$$

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 tree leaf.

Interpreting the formulas in terms of variation in time, equation 3 averages the features over time, which are related as they fall in the same mixture component. Equation 4 models the variation of related features over the entire video sequence, capturing subtle visual changes. Different mixture components will capture drastic variations in time, such as video shot changes.

Globally, the overall approach (with Random Forests) can be interpreted as a hard assignment FK approach. However, it differs from the standard FK approach in the following respects: (i) uses a fast Random Forest classification approach instead of the Gaussian Mixture Models; (ii) performs a hard assignment strategy, rather than a soft assignment, which is effective in the context of Random Forests.

3.3. Classification

The final component of the framework is the actual classifier that is fed with the FK-VLAD descriptors. Following the standard best practice [1], we use a SVM classifier. In the case of multimodal descriptors, each modality is fed to individual SVMs and their output confidence levels are (late) fused with a linear weighted combination:

$$CombMean(d, q) = \sum_{i=1}^{N} \alpha_i \cdot cv_i \tag{5}$$

where cv_i is the confidence value of classifier i for class q, $q \in \{1, ..., C\}$ with C 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 (see Section 4).

4. EXPERIMENTAL RESULTS

Experimental validation was conducted on several standard datasets, namely: *Blip10000* [12] — 15,000 video sequences

Table 1: Comparison to baselines on the *Blip10000* [12] dataset, MAP (best results are in bold).

Approach	*HoG*	*CN*
frame statistics, SVM RBF	0.182	0.223
Bag-of-Words, SVM RBF	0.232	0.263
VLAD, SVM RBF	0.254	0.314
proposed with k-means, SVM RBF	0.245	0.316
proposed, SVM RBF	**0.295**	**0.385**

Table 2: Comparison with the 2012 MediaEval results on *Blip10000* [12] dataset (best results are in bold).

Feature & approach	*MAP*
block-based audio features, 5-nearest neighbor [20]	0.192
SAD, proposed	0.472
color, texture and rgbSIFT descriptors [21]	0.350
HoG and CN, proposed	0.453
HoG, CN and SAD, proposed	**0.533**
text, Bag-of-Words [22]	0.523

from blip.tv with associated social metadata and automatic speech transcripts, labeled according to 26 video genres, e.g., "art", "gaming", "movies and television", etc; *VSD2013* [14] — 25 Hollywood movies annotated at shot level for their violence content (in total 43,923 shots; yes/no annotations); *UCF101* [13] — 13,320 realistic videos from YouTube labeled according to 101 action categories, e.g., "human-object interaction", "playing musical instruments", "sports".

Performance is assessed with two standard metrics: *accuracy* (the number of items correctly classified) and *MAP* (mean average precision).

Methods' parameter optimization is carried out on training data which is split in two fixed, equally sized parts, one for training parameters and the other for testing. Actual testing is conducted by adopting the optimal configuration.

For content description, we experimented with some standard approaches, known to perform well in many benchmarking scenarios [1]: *visual descriptors:* we compute Histograms of oriented Gradients (HoG — 81 values) [15] and color naming histograms (CN, a perceptually based color descriptor that maps colors into 11 universal color names — 11 values) [16]; *motion descriptors:* we compute Histograms of optical Flow (HoF — 72 values) [17]; *audio descriptors:* we compute some general-purpose audio descriptors, 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 (SAD — 196 values) [19].

Our goal is not to find the best descriptor combination but we want to show that the proposed framework is effective enough to allow state-of-the-art performance even when using standard descriptors. Depending on the dataset, we use different descriptor combinations. To reduce resulting Random Forest feature noise, final descriptor is decorrelated and reduced with Principal Component Analysis (PCA).

4.1. Video genre classification

A first validation experiment consisted in classifying video genres using the *Blip10000* [12] dataset. To compare with state-of-the-art results, we adopt the official dataset's evaluation scenario where training is performed on 5,288 videos and the actual evaluation on the remaining 9,550 videos.

Parameter tuning. Parameters are optimized on the training data. We determine that for this task best configuration is with PCA for HoG and SAD (20% reduction), no PCA for CN, SVM with non-linear RBF kernel, L2 with power normalization for the modified VLAD features and 100 random trees which is a good tradeoff for performance-feature size.

Comparison with baselines. We first compare our approach against several standard baseline approaches: Bag-of-Words, VLAD and simple frame statistics (use of feature mean and dispersion over all the frames). We also compare to the proposed approach when using standard k-means instead of the Random Forests for word assignments. Results are summarized in Table 1. The proposed description framework allows for an average improvement of more than 0.05 MAP over the best baseline and a similar improvement over the use of k-means instead of Random Forests.

Comparison to state-of-the-art. We compare with the best results reported at the 2012 MediaEval benchmarking on this dataset. Results are presented in Table 2. The proposed description framework achieves better performance than the use of highly semantic text descriptors (use of speech transcripts and metadata). Over the visual and audio features the improvement is significant, more than 0.18 MAP.

4.2. Violent scenes classification

A second validation experiment is for the classification of violent content in movies using the *VSD2013* [14] dataset. In particular, for this data, two different violence annotations were available, an objective one (physical violence) and a subjective one (violent content not suitable for a 8 year child). Both scenarios were tested. We adopt the dataset's official evaluation: 18 movies are used for training (32,678 shots) and the remaining 7 (11,245 shots) for evaluation.

Parameter tuning. Parameters are optimized on the training data. We determine that for this task best configuration is with PCA for HoG and CN (10% reduction), no PCA for SAD, SVM with non-linear RBF kernel, L2 with power normalization for the modified VLAD features and 12 random forests.

Comparison to state-of-the-art. We compare with the best results reported at the 2013 MediaEval benchmarking on this dataset. Results are presented in Table 3. The most efficient approaches remain those that include multimodal in-

Table 3: Comparison with the 2013 MediaEval results on *VSD2013* [14] dataset (best results are in bold).

Feature & approach	*MAP*
Objective violence annotation	
aural & visual descriptors, Multi Layer Perceptron [25]	0.3504
temporal, audio & visual descriptors, Multiple Kernel Learning [24]	0.4202
HoG, proposed	0.5601
SAD, proposed	0.6137
CN, proposed	0.6695
HoG, CN and SAD, proposed	**0.7202**
Subjective violence annotations	
temporal, audio and visual descriptors, Bayesian networks [23]	0.4479
HoG, proposed	0.7206
SAD, proposed	0.6276
CN, proposed	0.7206
HoG, CN and SAD, proposed	**0.7612**

formation. The proposed approach achieves best results in both scenarios. For the objective annotation we improve the state-of-the-art by more than 0.3 MAP while for the subjective annotations by 0.32 MAP.

4.3. Human action classification

The final experiment is for the classification of human actions on the *UCF101* [13] dataset. As for the previous experiments, we adopt the dataset's standard evaluation framework, where a quarter of the dataset (3,207 videos) are used for training and the whole dataset (13,320 clips) for the evaluation.

Parameter tuning. Parameters are optimized on the training data. We determine that for this task best configuration is with PCA for HoG (10% reduction), no PCA for CN and HoF, SVM with non-linear RBF kernel, L2 with power normalization for the modified VLAD features and 10 random trees for CN and HoG and 150 random trees for the HoF.

Comparison to State-of-the-Art. We compare with several state-of-the-art approaches from the literature. Results are presented in Table 4. The proposed approach achieved the second best accuracy, 74.1%. The highest accuracy, 87.90%, is obtained with discriminatively trained deep Convolutional Networks [30]. Although not more effective than the use of deep learning, the advantage of the proposed framework is in the use of simple global features, compared to the computationally more expensive Space-Time Interest Points (trajectories) or highly complex deep learning architectures.

4.4. Computational complexity

To assess the computational requirements of the proposed description framework, we experimented with the *Blip10000* [12] dataset. We used a standard PC with 2.9 GHz Intel Xeon CPU and 24GB of RAM without parallelization. Experiments were run for optimal parameters (see Section 4.1). For the entire processing chain we achieve 239 ms per frame using HoG ($\sim$ 6 seconds for 1 second of video) and 79 ms per frame using CN ($\sim$ 2 seconds for 1 second of video), from which input/output operations last 30 ms per frame and VLAD computation only 12 ms per frame. Considering the fact that no hardware acceleration was used nor any parallelization this is a more than reasonable processing time.

Table 4: Comparison with state-of-the-art on the *UCF101* [13] dataset (best results are in bold).

Approach	*Accuracy*
Cuboid descriptors [13]	43.90%
dense trajectories and VLAD [28]	52.10%
Convolutional Neural Networks [27]	65.40%
ordered trajectories and VLAD [29]	73.1%
CN, HoG and HoF, proposed	74.1%
ConvNet architecture [30]	**87.90**%

5. CONCLUSIONS

We proposed a new video representation framework that uses a fast word assignment approach by replacing standard k-means visual vocabulary assignment with a Random Forest approach. A modified version of Vector of Locally Aggregated Descriptor with Fisher Kernel representation is used for increasing the representative power of the descriptors and for capturing video temporal information. We demonstrated that our framework is highly general achieving state-of-the-art results on several classification scenarios and outperforming current state-of-the-art approaches, e.g., Bag-of-Words or use of highly semantic user-generated textual information. Future extension of this work will mainly address the adaptation to exploit the classification efficiency of deep learning networks.

6. ACKNOWLEDGEMENT

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.

7. 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", NIST, USA, 2013.

[2] 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.

[3] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition", IJCV, 103(1), pp. 60-79, 2013.

[4] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories", ICCV, pp. 3551-3558, 2013.

[5] F. Perronnin, J. Sanchez, T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification", ECCV, 2010.

[6] D. C. Ciresan, U. Meier, J. Masci, L. Maria Gambardella, J. Schmidhuber, "Flexible, High Performance Convolutional Neural Networks for Image Classification", IJCAI, 22(1), pp. 1238-1242, 2011.

[7] L. Breiman, "Random Forests", Machine Learning, 45(1), pp. 5-32, 2001.

[8] 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", Int. Journal of Multimedia Inf. Retrieval, online, 2014.

[9] H. Jégou, M. Douze, C. Schmid, P. Pérez, "Aggregating Local Descriptors into a Compact Image Representation", CVPR, 2010.

[10] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating Local Image Descriptors into Compact Codes", IEEE Trans. on PAMI, 34(9), pp. 1704-1716, 2012.

[11] C. Imre, J. Korner, "Information Theory: Coding Theorems for Discrete Memoryless Systems", Cambridge University Press, 2011.

[12] 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 MMSys, 2013.

[13] 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.

[14] 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", MTAP, 2013

[15] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes, "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection", IEEE Int. Conf. on Intelligent Transportation Systems, pp. 432-437, 2009.

[16] 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.

[17] B. Lucas, T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision", Imaging Understanding Workshop, 1981.

[18] Y. Yang, D. Ramanan, "Articulated Human Detection with Flexible Mixtures of Parts", IEEE Trans. on PAMI, 35(12), pp. 2878-2890, 2013.

[19] B.Mathieu, S.Essid, T.Fillon, J.Prado, G.Richard: "YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software", ISMIR, pp. 441-446, 2010.

[20] 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", MediaEval workshop, 2012.

[21] T. Semela, M. Tapaswi, H. Ekenel, and R. Stiefelhagen, "KIT at Mediaeval 2012 - Content-based Genre Classification with Visual Cues, MediaEval workshop, 2012.

[22] S. Schmiedeke, P. Kelm, and T. Sikora, "TUB @ MediaEval 2012 Tagging Task: Feature Selection Methods for Bag-of-(visual)-words Approaches", MediaEval Workshop, 2012.

[23] C. Penet, C.-H. Demarty, G. Gravier, P. Gros, "Technicolor/INRIA Team at the MediaEval 2013 Violent Scenes Detection Task", MediaEval Workshop, 2013.

[24] S. Goto, T. Aoki, "TUDCL at MediaEval 2013 Violent Scenes Detection: Training with Multimodal Features by MKL", MediaEval Workshop, 2013.

[25] M. Sjöberg, J. Schlüter, B. Ionescu, M. Schedl, "FAR at MediaEval 2013 Violent Scenes Detection: Concept-based Violent Scenes Detection in Movies", MediaEval Workshop, 2013.

[26] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V.L. Quang, Y.-G. Jiang, "The MediaEval 2013 Affect Task: Violent Scenes Detection", MediaEval Workshop, 2013.

[27] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fe, "Large-scale video classification with convolutional neural networks", CVPR, 2014.

[28] M. Jain, H. Jegou, and P. Bouthemy, "Better Exploiting Motion for Better Action Recognition", CVPR, 2013.

[29] O. V. Murthy and R. Goecke, "Ordered Trajectories for Large Scale Human Action Recognition", ICCV, 2013.

[30] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos", CVPR, 2014.