DCT-based Features for Categorisation of Social Media in Compressed Domain

Sebastian Schmiedeke, Pascal Kelm, Thomas Sikora

Communication Systems Group
Technische Universität Berlin
Germany
{schmiedeke,kelm,sikora}@nue.tu-berlin.de

***Abstract*—These days the sharing of videos is very popular in social networks. Many of these social media websites such as Flickr, Facebook and YouTube allows the user to manually label their uploaded videos with textual information. However, the manually labelling for a large set of social media is still boring and error-prone. For this reason we present a algorithm for categorisation of videos in social media platforms without decoding them. The paper shows a data-driven approach which makes use of global and local features from the compressed domain and achieves a mean average precision of 0.2498 on the Blip10k dataset. In comparison with existing retrieval approaches at the MediaEval Tagging Task 2012 we will show the effectiveness and high accuracy relative to the state-of-the art solutions.**

***Index Terms*—Genre classification, bag of words, compressed domain features**

I. Introduction

The possibilities arising from new technologies (such as Web 2.0) facilitate significantly the production and dissemination of new content. Automatic classification of videos enables users to easier find the desired content by categorising them into semantic categories or genres. Manual annotation is laborious due to the huge amount of newly generated data.

Our contribution to that is a framework that is able to classify user-generated video sequences into several thematic topics without decoding the video stream. This omission of the decoding procedure results in a reduction of the processing time. For our investigations we use the Blip10k dataset [1] which was developed within the MediaEval 2012 benchmark [1]. This dataset consists of 14,838 videos gathered from blip.tv included with shot boundary information [2]. This shot segmentation was carried out automatically by a software implementation which is not necessarily perfect. The videos are exclusively divided into 26 categories—chosen by its uploader—which are named in Figure 2 and cover a broad range of topics and styles. These videos' categories reflect rather a thematic topic than a genre; therefore an allocation to a specific category only based on visual features is a very hard task.

This work is an extension to our MediaEval 2012 Tagging Task [3] participation, there we classified these videos into these categories using bags of features derived from visual content and associated textual metadata, but here we focus only on visual features extractable from the compressed domain. Therefore, we analysed 158,446 key frames extracted from videos of the development set, the test set has 261,418 key frames respectively. The results of the subsequent classification of proposed features are then compared against each other and other approaches using pixel domain features.

As global features extracted from the compressed domain, we apply Colour and Edge Directivity descriptor (CEDD) [4] and Tiny Image descriptor described in [5], but using the CIELAB colour space on the reconstructed mini images. The discrete cosine transform (DCT) coefficients of each key frames are used as local features. The key frames are beforehand scaled in compressed domain to the same size.

This paper is structured as follows. In the next section, we cover the related work. We introduce our approach using features extracted from the compressed domain in section III. The results are shown in section IV and we finish with a conclusion summarizing our main findings.

II. Related Work

In recent works, a common approach for categorization is to employ global features to represent genres and categories. Ionescu et al. [6] applied global visual features to recognize video genres in broadcast material. Their multi-modal approach—audio, temporal and contour features were also used—achieved a very high accuracy on a database of 91 hours of TV programmes. They notice that heterogeneous content in certain genre led to lower accuracies than for content that have repetitiveness in their structure. In their approach binary classifiers were trained for each single genre, and then combined to a multi-genre decision using the majority voting rule. They investigated three classification methods—namely support vector machine (SVM), nearest neighbour (NN) and linear discriminant analysis—and concluded that the use of SVM lead to the best results.

Ekenel et al. [7] tested their approach with TV material, but also with user-generated content on YouTube and concluded that the results are not significantly influenced by the lower

[1] http://www.multimediaeval.org/mediaeval2012/

MMSP'13, Sept. 30 - Oct. 2, 2013, Pula (Sardinia), Italy.
978-1-4799-0125-8/13/$31.00 ©2013 IEEE.

image quality. Their multi-modal framework that used audio-visual cues as well as cognitive and structural information achieved a high classification accuracy (CA) of 92% for web videos with typical broadcast genres. The authors used a set of binary classifier to obtain multi-genre decisions. A support vector machines with RBF kernel was trained per each genre and feature, and finally the decisions were combined.

These previous works used features extracted from the pixel domain. Girgensohn and Foote [8] were the first who apply principal component analysis (PCA) with transform coefficients on shot frames for video genre classification. Wang et al. [9] surveyed various visual descriptors extractable from a compressed MPEG-1/2 stream. We showed in [10] the effectiveness of such features in the domain of classifying broadcast video streams. Web videos are typically not encoded in MPEG-1/2 but in H.264. Since our web videos are encoded in Ogg Theora, it is possible to adopt these descriptors to be able to extract feature from these types of web videos.

The use of visual words generated from compressed domain local features is presented in Sui et al. [11]. They showed an approach that reconstructed mini images by partly applying inverse DCT to the first few DCT coefficients of the luminance colour channel. On these mini images, scale-invariant feature transform (SIFT) features had been extracted and then clustered. The results of their binary image classification is not necessary worse than classifications of fully decoded images.

In contrast to Sui et al., we extract features directly from DCT coefficients. In order to obtain scale-invariant words, arbitrary resizing of intra-coded images in DCT space is required which was recently presented by Mukherjee and Mitra [12].

III. Methodology

Our proposed framework extracts visual features from the compressed domain of shared media. The provided video sequences of this dataset are encoded in Ogg Theora codec [2]. This codec has two types of coded frames (intra and inter coded). In order to reduce the temporal dimensionality of the video sequences we use intra-coded frames temporally close to key frames determined by the shot boundary detection. These frames are used to generate mini images on which global colour features are extracted. The local features are extracted from DCT coefficients of the Y colour channel at a regular grid using five different approaches. The characteristics of the video sequences are then presented using the bag-of-words (BoW) approach for local features. In contrast, the global features are directly fed into the subsequent classifier.

The flowchart of our approach is shown in Figure 1 and is described more precisely in the next subsections.

A. Global visual features

Intra coded frames carries all the image information which can be easily restored using the inverse DCT, but this processing step can be omitted if coarser resolutions are sufficient enough. An example where is coarse image resolution is sufficient is the Tiny Image descriptor that anyhow downscales images to 32×32 pixels. The Tiny Image descriptor represents the colour information of each pixel as feature vector, typically the RGB colour space is used, but we employ the more perceptual-adapted colour space CIELAB. The intra frame consists of multiple 8×8 DCT blocks for each colour channel, whereas each 0^{th} (DC) coefficient carries the average colour information for the respective 8×8 pixel segment. Thus, mini images can be generated by treating each 0^{th} DCT coefficient as the colour value for a pixel—after a proper range adoption. So, the resolution of the mini images is one-eighth of its original size. We extract global visual features, namely Tiny Image [5] and Colour and Edge Directivity Descriptor [4] from these reconstructed mini images. In previous work [13], we already compared the effectiveness of several global descriptors, therefore we choose for these two.

B. Visual Words Generation

We not only focus on global features but also on densely sampled DCT-based local features to enable our approach presented in [14] to work on this much larger dataset. In order to obtain square images tiles with a length of 5 % of the image width, the number of DCT blocks of each frame is proportionally reduced to 20 horizontal blocks. From each block a description is extracted as described in section III-B2. Once the local feature vectors are extracted, the visual word generation is performed by vector quantisation of decorrelated features.

1) Frame Resizing without Decoding: In order to obtain scale-invariant densely sampled visual words, the intra-coded key frame is proportional scaled to a fixed size by using a method initially proposed by Jiang and Feng [15] which was then refined by Mukherjee and Mitra [12]. Whereby the scaling is performed in DCT space, so the frame need not to be decoded beforehand. The basic idea is to recompose or to decompose spatial adjoining DCT blocks, and then performing a subband filtering of the coefficients. According to Mukherjee [12], image downscaling is equivalent to a recomposition of a block from $L \times L$ adjacent $N \times N$-point DCT blocks, where L is the downscaling factor:

$$\mathbf{D}^{LN \times LN} = \mathbf{A}_{L,N} \begin{vmatrix} \mathbf{D}_{0,0}^{N \times N} & \cdots & \mathbf{D}_{0,L-1}^{N \times N} \\ \vdots & \ddots & \vdots \\ \mathbf{D}_{L-1,0}^{N \times N} & \cdots & \mathbf{D}_{L-1,L-1}^{N \times N} \end{vmatrix} \mathbf{A}_{L,N}^{T}. \quad (1)$$

Subsequently, the $LN \times LN$-point DCT block $\mathbf{D}^{LN \times LN}$ needs to be subsampled with following equation:

$$D^{N \times N}(u,v) = \frac{D^{LN \times LN}(u,v)}{L}, \text{for } u,v = 0,\ldots,N-1. \quad (2)$$

Image upscaling operates similarly, here a $MN \times MN$-point DCT block is computed from a $N \times N$-point DCT block by inverting equation 2. This $MN \times MN$-point DCT block is

[2] http://www.theora.org/

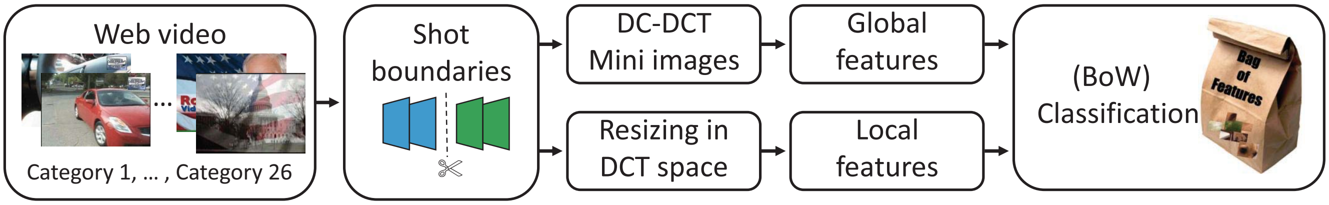

Fig. 1. Flowchart: Visual features are extracted from DCT coefficients derived from intra-codec key frames.

then decomposed into $M \times M$ blocks of size $N \times N$ each:

$$\begin{vmatrix} \mathbf{D}_{0,0}^{N\times N} & \cdots & \mathbf{D}_{0,M-1}^{N\times N} \\ \vdots & \ddots & \vdots \\ \mathbf{D}_{M-1,0}^{N\times N} & \cdots & \mathbf{D}_{M-1,M-1}^{N\times N} \end{vmatrix} = \mathbf{A}_{M,N}^{-1}\mathbf{D}^{MN\times MN}{\mathbf{A}_{M,N}^{-1}}^{T}, \tag{3}$$

where M is the upscaling factor.

Both scaling factors L and M are integer values, resizing with rational factors (e.g. $\frac{M}{L}$) is achieved by performing a upscaling with the integer M and then a downscaling by L, as described in [12]. The transformation matrix $\mathbf{A}_{K,N}$ is determined by

$$\mathbf{A}_{K,N} = \sqrt{\frac{2}{NK^2}}\left(\mathbf{B}_{K,N}\mathbf{C}_{K,N}^{T}\right). \tag{4}$$

The elements of $\mathbf{B}$ are calculated by

$$\begin{aligned} B(k,n) &= a \cdot \cos\frac{(2n+1)\cdot k\cdot\pi}{2KN} \\ &\text{for } k,n = 0,\ldots,KN-1, \\ &\text{where } a = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k=0 \\ 1 & \text{otherwise} \end{cases}. \end{aligned}$$

The elements of $\mathbf{C}$ are calculated by

$$\begin{aligned} C(k+m\cdot K, n+m\cdot K) &= a\cdot\cos\frac{(2\cdot n+1)\cdot\pi\cdot k}{KN} \\ &\text{for } k,n = 0,\ldots,N-1 \text{ and } m = 0,\ldots,K-1, \\ &\text{where } a = \begin{cases} 1 & \text{if } k=0 \\ \sqrt{2} & \text{otherwise} \end{cases}. \end{aligned}$$

These formulas guarantee arbitrary sizes of visual words, the description of those is described in the following section.

2) *Description using DCT Coefficients:* Each DCT block is now treated as a visual word, its description is generated by five different methods using the coefficients $\mathbf{D}$ of a DCT block of the Y colour channel, respectively its zigzag scanned version $\mathbf{d}$. We investigate different methods of generating descriptions since the different handling of DCT coefficients results in different charateristics. Visual word descriptions that do not consider the 0^{th} coefficient ($u = v = 0$) should be invariant against uniform luninance offset, while descriptions with absolute values of the coefficients should be invariant against its mirrored versions.

a) Quantised Coefficients: The feature vector is created by quantising the DCT coefficients $D(u,v)$ using the standard JPEG quantisation matrix $\mathbf{Q}$ [16]:

$$\mathbf{Q} = \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}. \tag{5}$$

The quantisation is then

$$D_q(u,v) = Q(u,v)\cdot\text{round}\left(\frac{|D(u,v)|}{Q(u,v)}\right). \tag{6}$$

The resulting feature vector $\mathbf{f}_a$ consists of the zigzag scanned and quantised absolute values of coefficients $\mathbf{d}_q$. So this feature vector is invariant to mirrored versions of the respective image section.

b) AC Coefficients: The feature vector created with this method simply consists of the zigzag scanned absolute values AC coefficients of a DCT block $\mathbf{d}$, therefore it is invariant against luninance offsets:

$$f_b(p) = |d(p+1)|, \text{for } p = 0,\ldots,62. \tag{7}$$

c) Logarithmic AC Coefficients: Feature vector $\mathbf{f}_c$ is created as proposed by Sim et al. [17] but applied on a single DCT block only. Sim et al. [17] report good retrieval results applying this method. Here the first zigzag scanned 51 AC coefficients are used to generated the description:

$$f_c(p) = \begin{cases} \log(|d(p+1)|) & \text{if } d(p+1) > 0 \\ 0 & \text{otherwise} \end{cases}, \text{for } p = 0,\ldots,50. \tag{8}$$

d) Pairwise averaged Coefficients: This feature vector is created by pairwise averaging each element of the upper triangular matrix of $\mathbf{D}$ with its' transposed element, therefore the feature is invariant against rotation by multiple of 90:

$$\begin{aligned} f_d(p) = 0.5\cdot\left(D_q(u,v) + D_q(v,u)\right), &\text{for } u,v = 0,\ldots,7 \\ &\text{and } p = u + 8v - 0.5(v^2+v). \end{aligned} \tag{9}$$

e) Diagonalised averaged Coefficients: Feature vector $\mathbf{f}_e$ is here created by averaging all elements of $\mathbf{D}$ lying on a diagonal according to the zigzag scanning scheme.

$$f_e(p) = \frac{1}{v_{p+1} - v_p} \sum_{i=v_p}^{v_{p+1}-1} f_a(i), \text{for } p = 0, \ldots, 14$$
$$\mathbf{v} = \{0, 1, 3, 6, 10, 15, 21, 28, 36, 43, 49, 54, 58, 61, 63, 64\}\,. \tag{10}$$

3) Vector Quantisation: The visual vocabulary V_{vis} is built from quantised feature vectors, since clustering is a time consuming step in BoW approaches. We introduce an approach for generating BoW histograms which does not require any clustering procedure. The straight forward method of vector quantisation is problematic, since the number of cells growths exponential with the dimensionality of the feature. A uniform quantisation of a descriptor with d dimensions would lead to q_l^d code words, where q_l is the number of quantisation levels. We tackle that problem by decorrelating the feature vectors and then quantifying each dimension separately. Therefore, we transform the features into a feature space with zero mean and a variance of one. This transformation is performed using principal component analysis of the covariance matrix $\mathbf{C} = \mathbf{F} \cdot \mathbf{F}^T$ for the corresponding features $\mathbf{F}$:

$$\mathbf{\Lambda} = \mathbf{V}^{-1}\mathbf{C}\mathbf{V}, \tag{11}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues λ and $\mathbf{V}$ as corresponding eigenvectors. Each single description F_i is then transformed by $F_{i,pca} = \mathbf{V}^T \cdot F_i$. Each dimension is in a different scale, depending on their impact on the total variance. In order to obtain features with unit variance, the transformed descriptor is scaled by $\frac{1}{\sqrt{\lambda_j}}$, where j is the index of the dimension. Divisions by zero are avoided by adding a small term $\epsilon = 10^{-6}$ to each eigenvalue λ.

$$\mathbf{F}_{white} = \mathbf{V}^T\mathbf{F}\left(\mathbf{\Lambda} + \epsilon\,\mathbf{I}\right)^{-1} \tag{12}$$

The result of this transformation is a feature space with zero mean, unit variance and decorrelated dimensions, since the covariance matrix is set to the identity matrix $\mathbf{I}$. Now, the features can be quantified using the 3σ rule, so the range to be quantified is set to $[-3; 3]$. For a uniform quantization, the quantization step is subsequently to $q_s = \frac{6}{q_l}$, where q_l is the desired quantization levels for each dimension. The amount of codewords depends on the quantization level q_l and dimensionality d of the feature: $|v_{vq}| = q_l \cdot d$. Each visual word of a feature $\mathbf{F}_i$ is determined by $cb_j = q_l \cdot \left(j + \frac{\mathbf{F}_{white,i,j}+3}{6}\right)$, $\forall j \in 0, \ldots, d-1$. Since the feature dimensions are linear independent, each feature vector $\mathbf{F}_i \in \mathbb{R}^d$ generates d visual words.

C. Classification & Fusion

The BoW histograms are classified with a multi-class SVM with histogram intersection (HI) kernel and cost parameter $C = 1$. The classification into multiple genres is obtained using the *one-vs-one* strategy and the majority voting rule. The HI kernel is defined by

$$\kappa(\mathbf{x}, \mathbf{y}) = \sum_i \min(x(i), y(i)). \tag{13}$$

The feature vectors of the global features are whether classified with a SVM with linear kernel ($\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$) or using the k-nearest neighbour approach and k-d trees. The results achieved with a single descriptor are then combined in late fusion manner using the weighted voting rule. So, each classifier output is treated as a vote to a certain category which is weighted with its normalised average precision to that category.

IV. Experiments

This section contains the experiments to show the performance of different visual features extracted from compressed domain. These experiments are carried out on the Blip10k dataset [1] that comprises 5,288 videos (158,446 shots) in the development set and 9,550 videos (261,418 shots) in the test set. This dataset contains almost randomly chosen social videos belonging to one of the 26 categories depicted in Figure 2 without concerning the video content or correctly tagged videos. The scenario is to predict the user-chosen category on blip.tv, including the *default category* that means the uploader does not choose any category. This “real world” dataset does not have a balanced distribution of categories, i. e. half of the video within this dataset belongs only to one of the five categories: *default category, politics, technology, music & entertainment, or educational.* Based on this unbalanced dataset, the mean average precision (mAP) is chosen as measurement for evaluation, although each video is classified to a unique category. The use of classification accuracy as evaluation measurement is not informative enough to evaluate classifiers for this dataset. Since the distribution of categories is unbalanced, a classifier predicting only the default category would lead to a CA of 0.1623, but to a mAP of 0.0068. The evaluated results of our local descriptors quantised with different quantisation levels into visual words are shown in Table I.

TABLE I
Results of different local DCT-based features with different quantisation levels (in mAP). A SVM with HI kernel is used as classifier.

Local feature	**VQ16**	**VQ32**	**VQ64**
quantised coeff. ($\mathbf{f}_a$)	0.1213	0.1242	0.1307
AC coeff. ($\mathbf{f}_b$)	0.1058	0.0985	0.0920
log AC coeff. ($\mathbf{f}_c$)	0.1178	0.1212	0.1202
pairwise avg. coeff. ($\mathbf{f}_d$)	0.1172	0.1180	0.1181
diagonalised avg. coeff. ($\mathbf{f}_e$)	0.1035	0.1095	0.1199

These local descriptors $\mathbf{f}_{a-e}$ described in section III-B2 are vector quantised with the quantisation levels $q_l = 16, 32, 64$ and a BoW histogram is generated from the resulting visual words for each video sequence. These BoW histograms are then classified using a set of SVM with histogram intersection kernels. As shown in Table I, the best result is achieved using the f_a local descriptor that consists of all DCT coefficients quantised with the standard JPEG quantisation matrix. A mean average precision of 0.1307 is obtained by visual words generated by vector quantisation with $q_l = 64$. The different local descriptors have different trends for scaling with the quantization level. In general, finer-grained vector quantisation lead to more discriminate feature spaces and result in higher precisions.

In the following, the global visual descriptors are described that extract their features from the compressed domain.

$TinyCEDD$: The Colour and Edge Directivity descriptor extracts its features from reconstructed mini images converted into RGB colour space. Each feature vector is separately classified using k nearest neighbour ($k = 6$) algorithm and a k-d tree. A decision for the whole video sequence is then obtained using consensus voting rule.

$TinyLab$ is the Tiny Image descriptor. This feature is extracted from the reconstructed mini image downscaled to 32×32 pixels and converted into CIELAB colour space. The feature vector is reduced using principal component analysis to 900 dimensions, and then classified using SVM with linear kernel. Here the features vectors of each video sequence are averaged beforehand.

TABLE II
RESULTS ACHIEVED WITH GLOBAL COMPRESSED DOMAIN FEATURE AND THEIR FUSION WITH THE LOCAL ONES.

Feature	Classifier	mAP
TinyCEDD	k-Nearest Neighbour ($k = 6$)	0.1466
TinyLAB	SVM, linear Kernel	0.1124
Fusion$_{wv}$	weighted voting	0.2498

Table II shows the results of these global visual descriptor. Among these both feature, the tiny version of CEDD achieves the higher mean average precision. All results can be also compared to the lower bound mAP achieved by random guessing ($mAP = 0.0026$) or choosing the predominant default category ($mAP = 0.0068$).

The average precision for each category of each single descriptors mentioned above is shown in Figure 2. It is shown that classifiers perform differently for the single categories, and therefore a combination of these classifier to a single system is reasonable. As depicted, the category *autos & vehicles* is distinguished best; an average precision of $AP = 0.7692$ is achieved. Whereas, the recognition of categories, such as *personal or auto-biographical*, *travel*, or *web development & sites*, is quite low—almost 0. The reason is seen in the visually indistinguishable content; *personal or auto-biographical* and *travel* are sub topics of *documentary*, while *web development & sites* videos look very similar to those belonging to *technology*.

Contrary to earlier investigations [14], global features achieves remarkable precisions compared to the local ones. A possible reason for relatively worse performance of these local features is the lack of rotation invariance in their current configuration. The result of our compressed domain features can be improved by combining them. In late fusion manner using the weighted voting rule, a mean average precision of 0.2498 is then achieved, labelled as "Fusion$_{wv}$" in Table II and in Figure 2. This figure depicts that this fusion achieves at least an AP of 0.1 for almost every single category. The best distinguished category remain *autos & vehicles*, same for the worst—*travel*. Finally, every third decision (36.5%) made on these 9,550 videos was right.

TABLE III
OUR FUSION RESULT IN COMPARISON TO OTHER APPROACHES USING THAT DATASET AND TWO BASELINES METHODS.

	Feature	mAP
our fusion	local+global visual	0.2498
UniCamp (from [1])	global visual	0.1238
ARF (from [1])	global visual, audio	0.1941
baseline 1	random	0.0026
baseline 2	default category	0.0068

Our fusion result achieved with global and local features extracted compressed domain is now compared to other approaches using that dataset. These results are listed in [1] and were achieved within the MediaEval 2012 Tagging Task. The purely data-driven approaches within this campaign came from the participants UniCamp and ARF. While UniCamp used a histogram of spatio-temporal motion patterns achieving a mAP of 0.1238, ARF used miscellaneous global colour descriptors resulting in a mAP of 0.1941. Compared to these results, our fusion result is remarkable considering that video sequences need not decoded for our approach. The results achieved in terms of mAP are shown in Table III.

V. CONCLUSION

The BoW approach proposed in this paper uses vector quantisation of decorrelated local features to avoid the time-consuming clustering procedure within generation of the visual words. The proposed approach only requires the compressed stream (but Huffman decoded) of video sequences from those the local and global visual features are extracted. The best proposed local descriptor uses all 64 coefficients quantised with the standard JPEG quantisation matrix of scaled DCT blocks. Our best result of $mAP = 0.2498$ is achieved by combining various classifier using local and global visual features. As already shown in a previous publication [14], the classification result on user-generated content can be significantly improved by including textual metadata.

In future work we will spend more effort in making our local descriptors extracted from DCT coefficients rotation invariant. We will also extent our BoW model by using hierarchically structured BoW histogram.

VI. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's FP7 under grant agreement number FP7-261743 (VideoSense).

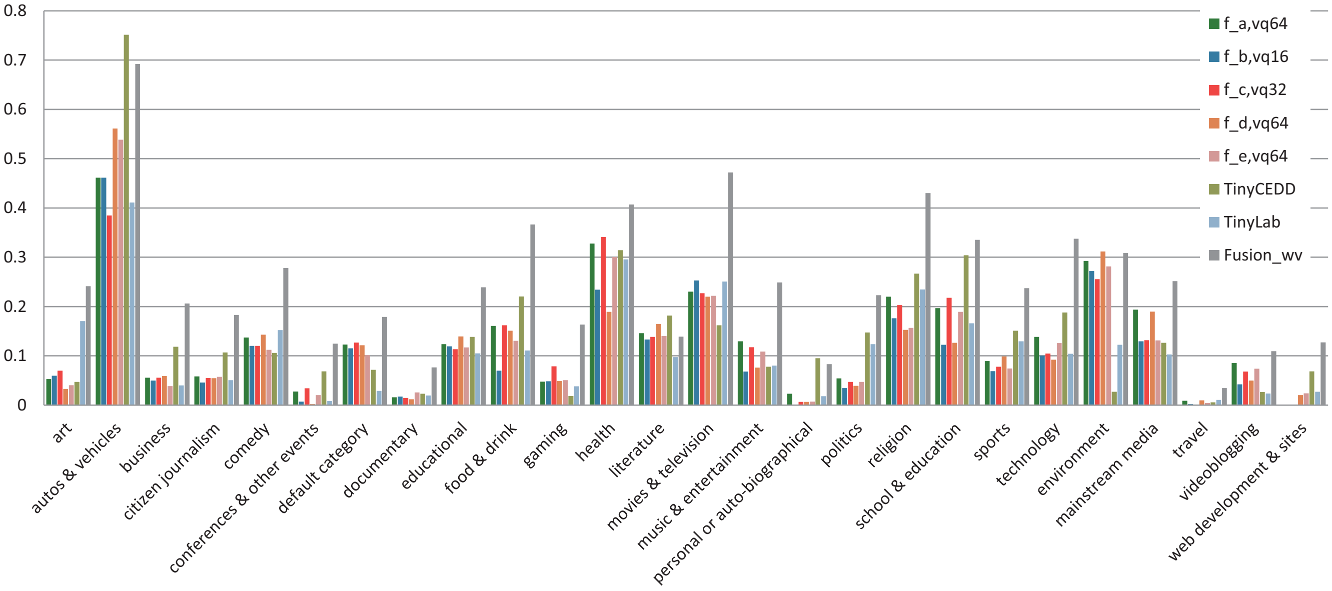

Fig. 2. Average precision for each category and each feature derived from the compressed domain and their fusion.

References

[1] S. Schmiedeke, P. Xu, I. Ferrané, M. Eskevich, C. Kofler, M. A. Larson, Y. Estève, L. Lamel, G. J. Jones, and T. Sikora, "Blip10000: A social Video Dataset containing SPUG Content for Tagging and Retrieval," in *ACM Multimedia System*. Oslo, Norway: ACM, February 26-March 1, 2013 2013.

[2] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-based video key frame extraction for low quality video sequences," in *10th Workshop on Image Analysis for Multimedia Interactive Services.*, 2009.

[3] S. Schmiedeke, C. Kofler, and I. Ferrané, "Overview of MediaEval 2012 Genre Tagging Task," in *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.

[4] S. Chatzichristofis and Y. Boutalis, "CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval," in *Computer Vision Systems*, ser. Lecture Notes in Computer Science, A. Gasteratos, M. Vincze, and J. Tsotsos, Eds. Springer Berlin / Heidelberg, 2008, vol. 5008, pp. 312–322.

[5] A. Torralba, R. Fergus, and W. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 11, pp. 1958 –1970, nov. 2008.

[6] B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan, and P. Lambert, "Content-based video description for automatic video genre categorization," in *The 18th International Conference on MultiMedia Modeling*, 4-6 January 2012, klagenfurt, Austria.

[7] H. K. Ekenel, T. Semela, and R. Stiefelhagen, "Content-based video genre classification using multiple cues," in *Proceedings of the 3rd international workshop on Automated information extraction in media production*, ser. AIEMPro '10. New York, NY, USA: ACM, 2010, pp. 21–26. [Online]. Available: http://doi.acm.org/10.1145/1877850.1877858

[8] A. Girgensohn and J. Foote, "Video classification using transform coefficients," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 6, 1999, pp. 3045–3048 vol.6.

[9] H. Wang, A. Divakaran, A. Vetro, S.-F. Chang, and H. Sun, "Survey of compressed-domain features used in audio-visual indexing and analysis," *Journal of Visual Communication and Image Representation*, vol. 14, no. 2, pp. 150 – 183, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320303000191

[10] R. Glasberg, S. Schmiedeke, M. Mocigemba, and T. Sikora, "New real-time approaches for video-genre-classification using high-level descriptors and a set of classifiers," in *Proc. IEEE International Conference on Semantic Computing*, 4–7 Aug. 2008, pp. 120–127.

[11] L. Sui, J. Zhang, L. Zhuo, and Y. Yang, "Research on pornographic images recognition method based on visual words in a compressed domain," *Image Processing, IET*, vol. 6, no. 1, pp. 87 –93, feb. 2012.

[12] J. Mukherjee and S. K. Mitra, "Arbitrary resizing of images in dct space," in *Vision, Image and Signal Processing, IEE Proceedings-*, vol. 152, no. 2. IET, 2005, pp. 155–164.

[13] P. Kelm, S. Schmiedeke, and T. Sikora, "Multimodal geo-tagging in social media websites using hierarchical spatial segmentation," in *Proceedings of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM SIGSPATIAL, Nov. 2012, p. 8.

[14] S. Schmiedeke, P. Kelm, and T. Sikora, "Cross-modal categorisation of user-generated video sequences," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, 2012, pp. 25:1–25:8. [Online]. Available: http://doi.acm.org/10.1145/2324796.2324828

[15] J. Jiang and G. Feng, "The spatial relationship of dct coefficients between a block and its sub-blocks," *Signal Processing, IEEE Transactions on*, vol. 50, no. 5, pp. 1160 –1169, may 2002.

[16] JPEG Group, "libjpeg. 6b ed." 1998. [Online]. Available: http://www.ijg.org/

[17] D.-G. Sim, H.-K. Kim, and R.-H. Park, "Fast texture description and retrieval of dct-based compressed images," *Electronics Letters*, vol. 37, no. 1, pp. 18–19, 2001.