H.264 COMPRESSED VIDEO CLASSIFICATION USING HISTOGRAM OF ORIENTED MOTION VECTORS (HOMV)

Sovan Biswas *R. Venkatesh Babu*

Video Analytics Laboratory
Supercomputer Education and Research Centre
Indian Institute of Science
Bangalore, India

ABSTRACT

In this paper, we have proposed a simple and effective approach to classify H.264 compressed videos, by capturing orientation information from the motion vectors. Our major contribution involves computing Histogram of Oriented Motion Vectors (HOMV) for overlapping hierarchical Space-Time cubes. The Space-Time cubes selected are partially overlapped. HOMV is found to be very effective to define the motion characteristics of these cubes. We then use Bag of Features (BOF) approach to define the video as histogram of HOMV keywords, obtained using k-means clustering. The video feature, thus computed, is found to be very effective in classifying videos. We demonstrate our results with experiments on two large publicly available video database.

Index Terms— Video Classification, Compressed Domain, H.264, Histogram of Oriented Motion Vectors, Bag of Features

1. INTRODUCTION

With ever growing number of videos on web, the problem of video classification has grabbed the attention of vision researchers all over the world. On an average, 72 hours of videos are added only to youtube itself per minute all across the world [1]. Considering, on an average, if each video is of 3 minutes, then around $72 \times 60/3 = 1440$ videos per minute and $1440 \times 60 \times 24 = 2073600$ videos are being added every day! With such large amount of data being uploaded on the web, the need for video classification becomes crucial. Thus, there is an urgent need for large scale video classification for efficient video retrieval, annotations, etc.

Videos can be classified based on three modalities namely visual, audio and text material associated with it [2]. In this paper, we mainly focus on video content based classification. Recently, significant developments have undergone in this domain, resulting in better and faster automatic video classification. But, almost all of them require decompressed video and pixel level processing [3, 4]. As videos are stored in one or other compressed format, it is intuitive to develop algorithms in compressed domain for faster analysis. If one starts decoding each video for analysis, then just decoding all videos, uploaded in a day, will only take more than a day! Hence, the apparent need to look for approaches that handle video classification on compressed videos. Though, compressed domain processing is faster, the issue becomes complex due to lack of cues except for motion vectors and other compression parameters.

Motion Vectors (MVs) are integral part of any video compression technique, including H.264, which is widely used compression standard to encode high definition videos, and are considered to be coarse approximation of optical flow. But in H.264/AVC [5], motion estimation is more precise than any previous compression standards because of the use of Variable block-size motion compensation and Quarter pel motion estimation technique. Variable block-size motion compensation supports motion prediction for 16×16 to 4×4 block sizes, enabling accurate motion prediction. The supported prediction block sizes include 16×16, 16×8, 8×16, 8×8, 8×4, 4×8, and 4×4 and can be used together in a single macro block. half pel and quarter pel motion prediction are then used to predict block motion with more accuracy, resulting in nearly optical flow like characteristics.

Along-with, MVs being integral part of compression, they depict same traits among intra category members and dissimilarity across categories. The objective of this article is to classify video, based on content, by finding similarity of MVs patterns within a category.

The subsequent portion of the paper is organized as follows. We start with related work in section 2. Section 3, describes the proposed algorithm followed by experiments on **HMDB51** and **UCF50** database and analysis of the results in Section 4. Section 5 concludes the paper.

2. RELATED WORK

Prior to this, a great deal of work has gone into video classification. According to the review of Brezeale et al. [2], one can classify videos using different modalities. Modalities used for video classification can broadly be defined into three major categories - text based classification, audio based classification and video content based classification.

Much of the research on video classification is based on multiple modalities. Huang et al. [6] in his proposition suggested the use of Hidden Markov Model (HMM) on multiple modalities rather than single modality. On the other hand, Wang et al. [7], suggested the use of hybrid approach which involves different models for different modalities like SVM for text and Gaussian Mixture Modal for audio-visual features.

Among video content based approaches, Dimitrova et al. [8] uses face and text to determine the content of the video and used it in HMM model to classify or retrieve videos. Wang et al. [3] proposes a combined model of holistic spatial layout and temporal motion pattern for video classification. In [4], importance of salient region detection to compute effective features for video classification is presented by Rapantzikosa et al. Chaudhry et al. [9] proposed Histogram of Optical Flow (HOOF) feature alongwith Binet Cauchy Kernel on non dynamical system for action recognition (in effect 'Action' being content of the video). HOMV feature computed is

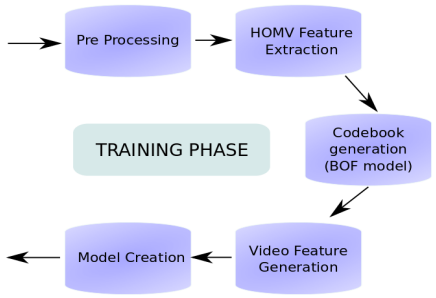
similar to HOOF feature only in terms of feature extraction, but differ drastically in other aspects such as region of interest, hierarchical spatio-temporal integration.

In compressed video analysis, majority of the research involves moving object segmentation in surveillance setup. Babu et al. [10], among the first few who used motion vector of compressed MPEG video for segmentation. More recently, Poppe et al.[11] and Verstockt et al.[12] introduced macro block size and macro block type of H.264 stream as new reliable parameters respectively. Moving to action recognition, Babu et al. [13] proposed MPEG MV based features along with HMM modeling and motion history information [14]. Motion Similarity based on H.263 MVs between consecutive frame was harnessed by Yeo et al.[15] to perform action recognition. But, almost all of the scenarios in action classification involves surveillance videos without any significant camera motion.

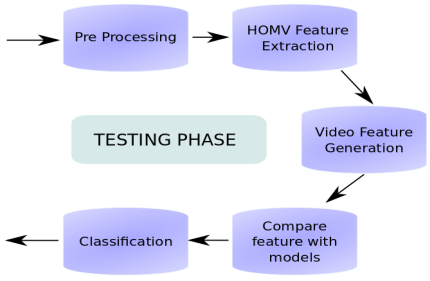
Even with large amount of research in recent times, video classification based on actions remains an open problem for researchers as the amount of data to be processed is still high for videos and thus requires more computing time.

3. PROPOSED ALGORITHM

The proposed algorithm has mainly three stages a) Preprocessing, b) HOMV feature extraction, and c) Video Feature Extraction. The block diagram (Fig.1) summarizes the training and testing module for video classification.



(a) Training module



(b)Testing module

Fig. 1. Block diagram of the proposed approach

3.1. Preprocessing

Preprocessing of the raw MVs from H.264/AVC involves removing the noisy motion vectors, estimating the camera parameters and finding the region of interest in the video.

Most of the noisy MVs have huge magnitude in the order of size of the video frame. Thus, MVs which are of length more than 10% of the frame size are truncated to zero, effectively neglecting noisy motions for future computations.

Secondly, video sequences belonging to same categories might have different camera motion that needs to be compensated. Camera parameters are estimated using eqn.(1), where s is the scale factor, p_3 and p_4 are the pan rate and tilt rate respectively. (x, y) and (x', y') being current and future locations of the blocks [16].

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = s \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} p_3 \\ p_4 \end{pmatrix} \tag{1}$$

MVs are compensated using the camera parameters estimated through eqn. (1).

The content of the video is the most essential cue for our objective and picking up the effective content (*viz.* region of interest) is very important aspect. We employ a simple region of interest (ROI) extraction based on spatial motion orientation gradient and motion magnitude gradient. The temporal accumulation of the above gradients are performed to achieve better ROI estimation. In effect, ROI is the region where MV changes frequently for a temporal bunch of frames. Mathematically,

$$ROI = \left(\sum_{i-k}^{i} (\nabla(M) + \nabla(O)) \right) > Th \tag{2}$$

where, M and O are Magnitude Image and Orientation Image with values normalized between 0 and 1. ∇ denotes image gradient. i is current frame and k being number of previous frames used (we have used k = 7). Though very naive, this provides better approximation of features to be extracted.

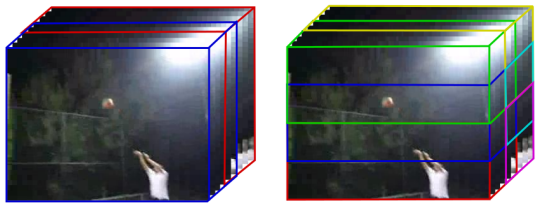
3.2. Feature extraction

Feature Extraction involves Space-Time cube generation, HOMV extraction and Orientation Normalization of the feature.

3.2.1. Space-Time Cubes Generation

Preprocessed MVs for a video are divided into temporal cubes from partially overlapping b frames. Each temporal cube is further divided into three levels of spatial cubes resulting in hierarchical Space-Time Cubes. At each level, the Space-Time cubes have same division, and across level, varies along spatially. We defined the cubes as $1 \times 1 \times 1$ at level I (coarse), $3 \times 1 \times 1$ at level II (medium) and $5 \times 1 \times 1$ at level III (finer). The finer cubes are formed by partially overlapping cubes along rows in order to retrieve interesting localized motion patterns. As the video could have interesting motion anywhere in the video, we need to generate features which are independent of the location. For example, a person swinging a baseball bat could be positioned at top half in one of the videos and bottom half in another. Generating features that are location independent, is crucial in this case, which we can achieve by dividing video into partially overlapped Space-Time cubes along rows. We do not divide the video along columns to handle left right symmetry. As observed in our experiments that over dividing a video, results in low spatial relation and reduces classification accuracy. Considering the above, we used HOMV creation

based on left-right symmetry. Figure 2 illustrates the spatial and temporal distribution of the motion vectors in Space-Time cubes.



(a) $1 \times 1 \times 1$ (b) $3 \times 1 \times 1$

Fig. 2. Space-Time cubes. a) $1 \times 1 \times 1$ is two Space-Time cubes with temporal overlap (Level I) b) $3 \times 1 \times 1$ is three Space-Time cubes with spatial overlap (Level II) [Images are best viewed in color]

3.2.2. *Region of Interest Features*

To emphasize more on the content, higher weightage is assigned to ROI. As coarser level motion feature captures the camera motion information, we ensured equal weight for all MVs and used them for feature computation. But, finer levels are intended to capture interesting localized motion, hence we formed the feature based only on MVs present in ROI.

3.2.3. *Histogram of Oriented Motion Vector (HOMV)*

We compute HOMV for each Space-Time cubes at each layer. Similar videos could have mirror properties ie., a person walking left to right is equivalent to person walking right to left (left-right symmetry). We define HOMV as histogram of motion vectors binned on primary angle and weighted according to its magnitude. Let the dimension of HOMV is n (number of orientation bins). Figure 3 further illustrates the orientation bins.

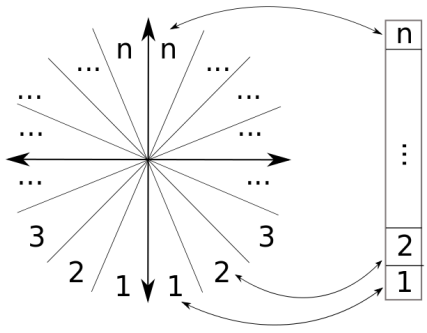


Fig. 3. Orientation Bins.

3.2.4. *Orientation Normalization*

All these HOMV features are direction normalized with respect to the coarse HOMV. This is achieved by wrapping around all the orientation bins with respect to the maxima of coarser HOMV bin.

Algorithm 1 ***HOMV Feature***

Input: *motion vectors for each Space-Time cubes.* **n** = number of orientations.
Output: *HOMV.*

MV = motion vector for Space-Time cubes.
orientation = $\lfloor tan^{-1}(MV_y/MV_x) * n/\pi \rfloor$.
magnitude = $\sqrt{(MV_x^2 + MV_y^2)}$.
initialize : feature $= \mathbf{0}_{1\times n}$

for all orientation at location (x, y) in MV **do**
feature(orientation(x, y)) = feature(orientation(x, y)) + magnitude(x, y)
end for

$HOMV = feature/\|feature\|_1$

3.3. Video feature

Given, the set of the HOMV features for each Space-Time cubes, we build a bag of features (BOF) for each hierarchical level. This requires building codebook for each hierarchy. In our experiments, we have randomly sampled HOMV features from each training video, forming a subset for each hierarchical level. We then optimized the subsets to form optimal codebook using k-means clustering.

Each Space-Time cube is then represented by the HOMV codebook based on the nearest word (Euclidean distance). We form histogram of words for each hierarchical level and then concatenate them to form the video level feature. We provided higher weight for finer level and lower weight for coarser levels.

$$F = [0.25 * f_{coarse}, 0.5 * f_{medium}, f_{fine}] \quad (3)$$

where, F is the video feature. And, f_{coarse}, f_{medium} and f_{fine} denotes BOF feature at each level.

4. EXPERIMENTS

In this section, we describe the datasets used for the evaluation as well as the evaluation procedure. We have conducted experiments on two large video databases to demonstrate the capability of our algorithm to handle wide range of variations with reasonable accuracy. Though both of them being action dataset we have used them as videos/action clips for classification based on content. Since, these datasets are not encoded in H.264 format, we encoded them in H.264 using Baseline profile with 1 reference frame. Group of Pictures (GOP) length is set to 30 and videos are encoded at a rate of 25 frame per sec. We have used libsvm (SVM classifier) for classification [17].

4.1. Dataset Used

UCF 50 is an action dataset having 50 actions [20, 19]. For experiments, we randomly selected 100 videos for each categories and divided them into 70:30 sets for training and testing. We ensured that the same video clip is not used for training and testing. We have also performed 3-fold cross validation.

Human Motion Database 51 (HMDB51) [18] is the one of the newest and most challenging database for action recognition. The evaluation criteria used for the dataset is same as that specified by [21]. We have used three distinct splits for training and testing. The

(a) UCF50 (b) HMDB

Fig. 4. Part of Dataset. courtesy [18] [19]

Approaches	UCF50	HMDB51
GIST[22]	38.8 (-)	13.4 (-)
HOG/HOF[23]	47.9 (-)	20.2 (-)
C2[21]	- (-)	23.2 (-)
Action Bank[24]	57.9 (12210 secs)	26.9 (-)
Proposed	**40.1 (2.7 secs)**	**18.3 (1.7 secs)**

Table 1. Results Comparisons : Accuracy (Execution Time per video)

splits are generated by randomly selecting 70 training and 30 testing clips by setting the constraint that each category follows 70:30 balance and no testing clip is used for training and vice versa.

We have compared the results with state of the art algorithms specified for these databases. Table 1 compares the results of recent algorithms and benchmarks. According to the best of our knowledge, all the benchmarked results documented in the table involves decompressed and pixel level processing. Our proposed compressed domain approach achieves comparable classification accuracy, with a high reduction in execution time, compared to pixel domain approaches. Figure 5 illustrates the confusion matrix for both dataset. UCF50 gives dominant diagonal for confusion matrix with highest classification rate of about 80%. HMDB gives a maximum classification rate of about 60%.

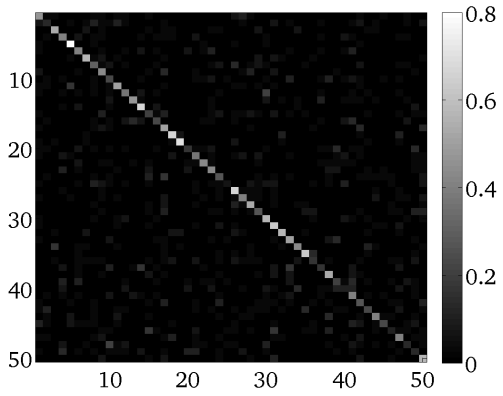
4.2. Analysis

To compare the effect of ROI on accuracy of classification, we simulated results with and without ROI computation on UCF 50 dataset. The results are shown in table 2. Clearly, the use of ROI provides better representation of video content.

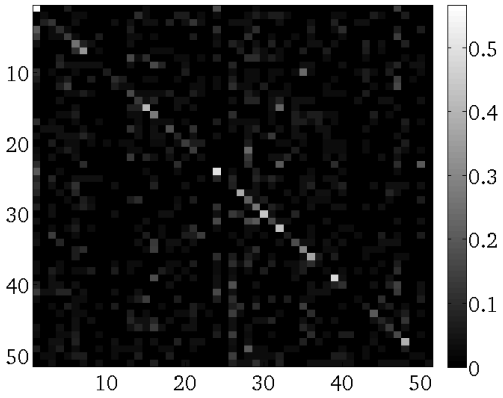
Motion compensation (MC) is introduced to compensate the camera motion. The results suggest only a slight improvement due to MC. This could be due to the fact that camera motion sometime provides important cue for classification. For example 'Military Parade' (UCF 50) will mostly associated with panning motion which help in better identification of the category (refer table 2).

Orientation normalization (ON) with respect to the highest global HOMV of the video is found to be important aspect. As depicted in the experiments, ON normalizes direction for HOMV features which helps in better classification (refer table 2).

Other parameters like number of keywords (K) and number of orientation bins (n), are empirically found to give good results for $K = 100$ and $n = 10$ respectively. But, the results are highly dependent on temporal size of Space-Time cubes. With increase in temporal size from $b = 6$ to 10 (see 3.2.1), the results drop from 40.1% to 36.9%. We have set $b = 6$ for all our experiments.



(a) UCF50



(b) HMDB

Fig. 5. Confusion Matrix for Datasets

	Without ROI	Without MC	Without ON	Actual
UCF50	31.9	38.8	34.67	40. 1

Table 2. Comparisons of effect of different parameters on UCF 50

5. CONCLUSION AND FUTURE WORK

We have proposed a compressed domain technique to classify H.264/AVC videos. The method mainly harness the fact that motions corresponding to similar content will follow similar orientation pattern. We have suggested effective approach to capture this pattern through oriented histograms of Space-Time cubes. A video is represented as combination of key oriented histograms which results in single feature vector for a single video. Experiments demonstrate comparable results to the state of the art even in compressed domain. Although initial experiments are encouraging, capabilities of HOMV are yet to be harnessed. The future work involves extending the capability of HOMV to other compressed video analysis problems. Also, other features of H.264 compressed videos can be harnessed along with it to achieve better classification rates.

6. ACKNOWLEDGEMENT

This work was supported by CARS (CARS-25) project from CAIR, DRDO, Govt. of India and VADS project from DST, Govt. of India.

7. REFERENCES

[1] "Youtube statistics," http://www.youtube.com/t/press_statistics.

[2] D. Brezeale and D.J. Cook, "Automatic video classification: A survey of the literature," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 3, pp. 416–430, 2008.

[3] Y. Wang and R. Gaborski, "Automatic video classification using holistic spatial features and optical flow," in *International Conference on Image Processing, Computer Vision, and Pattern Recognition*, 2011.

[4] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, and S. Kollias, "Spatiotemporal saliency for video classification," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 557–571, 2009.

[5] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h. 264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[6] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E.K. Wong, "Integration of multimodal features for video scene classification based on hmm," in *Multimedia Signal Processing*. IEEE, 1999.

[7] P. Wang, R. Cai, and S.Q. Yang, "A hybrid approach to news video classification multimodal features," in *Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003*. IEEE.

[8] N. Dimitrova, L. Agnihotri, and G. Wei, "Video classification based on hmm using text and faces," in *European Conference on Signal Processing*, 2000.

[9] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *IEEE Conference on Computer Vision and Pattern Recognition, 2009*. IEEE.

[10] R Venkatesh Babu, KR Ramakrishnan, and SH Srinivasan, "Video object segmentation: a compressed domain approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 4, pp. 462–474, 2004.

[11] Chris Poppe, Sarah De Bruyne, Tom Paridaens, Peter Lambert, and Rik Van de Walle, "Moving object detection in the h. 264/avc compressed domain for video surveillance applications," *Journal of Visual Communication and Image Representation*, vol. 20, no. 6, pp. 428–437, 2009.

[12] Steven Verstockt, Sarah De Bruyne, Chris Poppe, Peter Lambert, and Rik Van de Walle, "Multi-view object localization in h. 264/avc compressed domain," in *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009*. IEEE.

[13] R Venkatesh Babu, B Anantharaman, KR Ramakrishnan, and SH Srinivasan, "Compressed domain action classification using hmm," *Pattern Recognition Letters*, vol. 23, no. 10, pp. 1203–1213, 2002.

[14] R Venkatesh Babu and KR Ramakrishnan, "Recognition of human actions using motion history information extracted from the compressed video," *Image and Vision computing*, vol. 22, no. 8, pp. 597–607, 2004.

[15] Chuohao Yeo, Parvez Ahammad, Kannan Ramchandran, and S Shankar Sastry, "High-speed action recognition and localization in compressed domain videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1006–1015, 2008.

[16] Y.P. Tan, D.D. Saur, S.R. Kulkami, and P.J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 1, pp. 133–146, 2000.

[17] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[18] "Hmdb51 dataset," http://serre-lab.clps.brown.edu/resources/HMDB/.

[19] "Ucf dataset," http://crcv.ucf.edu/data/UCF50.php#Results_on_UCF50.

[20] Kishore K. Reddy and Mubarak Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, pp. 1–11, 2012.

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

[22] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[23] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al., "Evaluation of local spatio-temporal features for action recognition," in *BMVC-British Machine Vision Conference*, 2009.

[24] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.