ICDI International Conference on Disruptive Innovation 2016

Compressed Domain Feature Analysis for Efficient Action Recognition

Zar Zar Tun
The University of Information Technology (UIT)
Yangon, Myanmar
zarzarhtun@uit.edu.mm

Khin Htar Nwe
The University of Information Technology (UIT)
Yangon, Myanmar
khinhtarnwe@uit.edu.mm

ABSTRACT

Video surveillance system that use video cameras to transmit a signal to a specific place, on a limited set of monitors by Closed-circuit television (CCTV) that may need monitoring such as banks, casinos, airports, military installations, and convenience stores. Even though CCTV systems may monitor a particular event, it still requires recognized data are to be more accurate and appearance invariant when abnormal activities are occurred and analysed. Automatic Teller Machines (ATM) are widely used over the world for automatic banking transactions. Some of ATM users are facing with unlawful activities such as money snatching and customer assault by murder. CCTV devices are set up for security that becomes an important factor for covering criminal activities. Many researchers are trying to solve this issue in different ways of Human Activity Recognition.

In our system, we tend to use compressed video sequences MPEG flow instead of computationally expensive Optical Flow (OF). Compressed domain features such as Motion Vectors (MV) and Discrete Cosine Transform (DCT) coefficients can be achieved by partial decoding that needs no prior segmentation and no spatial or temporal assignment therefore it can reduce time requirement and can effect in minimal training. Mapping of compressed domain features are effective for no variation in actions recognition because of noise reduction in motion contents. Bag-Of-Words model based descriptors are created for each group of features by motion segmentation getting accurate feature description. Our system tends to create reliable and real time action recognition with compressed domain features analysis.

Keywords: Action Recognition, Motion Vectors, DCT coefficients, Noise Reduction, Bag-of-Words model, Motion Segmentation.

1 INTRODUCTION

Human action recognition based on computer vision is a significant role in many applications, such as intelligent video surveillance, video content analysis, and video retrieval. The widespread operational uses of closed circuit television or video surveillance as a security measure because it is cost-effective and much safer way of providing a crime. That can provide users with valuable benefits such as digital storage and remote accessibility, but there is a drawback to retain image quality for pattern recognizer for the large variations of human appearances, posture and body size into the video system. ATM transactions are quick and convenient, but the machines and the areas surrounding them can be susceptible to criminal activity if not properly protected. Using CCTV camera in ATM camps becomes threaten because of the requirement of CCTV system mainly in pattern recognizer for complexity and inaccurate recognition in real time video processing.

Human activity recognition becomes an increasing research trend and many researchers have developed various Optical Flow methods tend to enhance the quality of image appearance but it is

still computationally very expensive and resulting inaccurate action recognition. We tend to use Compressed Domain based MPEG flow features for efficient action recognition. Compressed Domain features such as the availability of motion vectors and pixel values in coded forms called DCT coefficients can indirectly provide motion and intensity information for action detection by avoiding the need to re-perform motion estimation. That can reduce required large amount of estimation time to get the exact motion vectors related with the whole image or specific parts such as rectangular blocks, arbitrary shaped patches or even per pixel. DCT coefficients that correspond to transformed residue data represent information, the block based motion vectors fail to capture.

Motion Vectors can represent both real object movement and fake object movement that sometimes introduces large amount of noise, lead to inefficient accuracy. Therefore our system made noise reduction by filtering methods to get reliable outcome. There is no motion information in I frame that can store only DCT information and can give texture information needed to flag motion reliability. Our system tends to map these specific features into related separate features clusters to use in feature description by motion feature segmentation. Separate features groups are created into related Bag-of-Words models that are histograms of a vector of occurrence counts of a vocabulary of local image features efficient for action classification and recognition.

The main goal of this paper is to take advantages of compressed domain information that are than classified and mapped into related bag-of-words model. Processing the compressed domain features reduce the amount of effort involved in full decompression, keeps the storage cost low and can improve the speed of feature extraction getting more accurate features with low computation complexity. This remainder of paper is organized as follows. Section 2 provides about the related works. Section 3 describes about proposed system. Section 4 describes about motion feature analysis and Section 5 presents about simulation experiments. Section 6 draws the conclusion and further extension.

2 LITERATURE REVIEW

In paper [1], authors proposed algorithm utilizes cues from quantization parameters and motion vectors extracted from the compressed video sequences for feature extraction and further classification using Support Vector Machines (SVM). In paper [2], authors tackled the recognition problem by replacing optical flow with motion vectors from the compressed domain. Authors proposed a set of residue boundary histograms (RBH) features for action recognition. In paper [3], authors introduced the notion of quantifying the motion involved, through Motion Flow History (MFH). The encoded motion information readily available in the compressed MPEG stream is used to construct the coarse Motion History Image (MHI) and the corresponding MFH.

In paper [4], authors proposed a novel motion correlation measure based on computing motion similarity using compressed domain features which can be extracted with low complexity. In paper [5], authors considered the problem of improving the optical flow field in MPEG sequences. This paper address the issue of robust, incremental, dense optical flow estimation by combining information from two different velocity fields: the available MPEG field and the one inferred by a multi resolution robust regularization technique applied on the DC coefficients. In paper [6], authors avoid motion estimation and design fast descriptors using motion information from video compression. Then follow Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH) for local descriptors for efficient action recognition.

In paper [7], authors proposed several speed-ups for densely sampled HOG, HOF and MBH descriptors. Then, it investigated the trade-off between accuracy and computational efficiency of descriptors in terms of frame sampling rate and type of Optical Flow method and the trade-off between accuracy and computational efficiency for computing the feature vocabulary, using and

comparing most of the commonly adopted vector quantization techniques. In paper [8], authors proposed framework that is based on the combination of static posture information and rotational features which are extracted from the Binary Silhouettes acquired using the texture based segmentation approach to develop a novel descriptor. In paper [9], authors combined the results of colour segmentation which is effected by the precision of motion segmentation and Motion estimation getting very useful generic automatic segmentation algorithm for all kinds of video sequences. In paper [10], authors proposed a video based framework that efficiently identifies abnormal activities happening at the ATM installations and generates an alarm during any untoward incidence. The proposed approach makes use of Motion History Image (MHI) and Hu moments to extract relevant features from video.

3 OVERVIEW OF THE SYSTEM

Feature extraction absolutely efficient for action recognition using traditional action recognition approaches are too slow for real-time or large scale applications. Lower speed features extraction can cause higher complexity and prevents from scaling up to real size problem getting invariant results. Our proposed system tends to detect all occurrences of query video sequence in a test video, and thereby recognizing an action as taking place at some specific time and invariant.

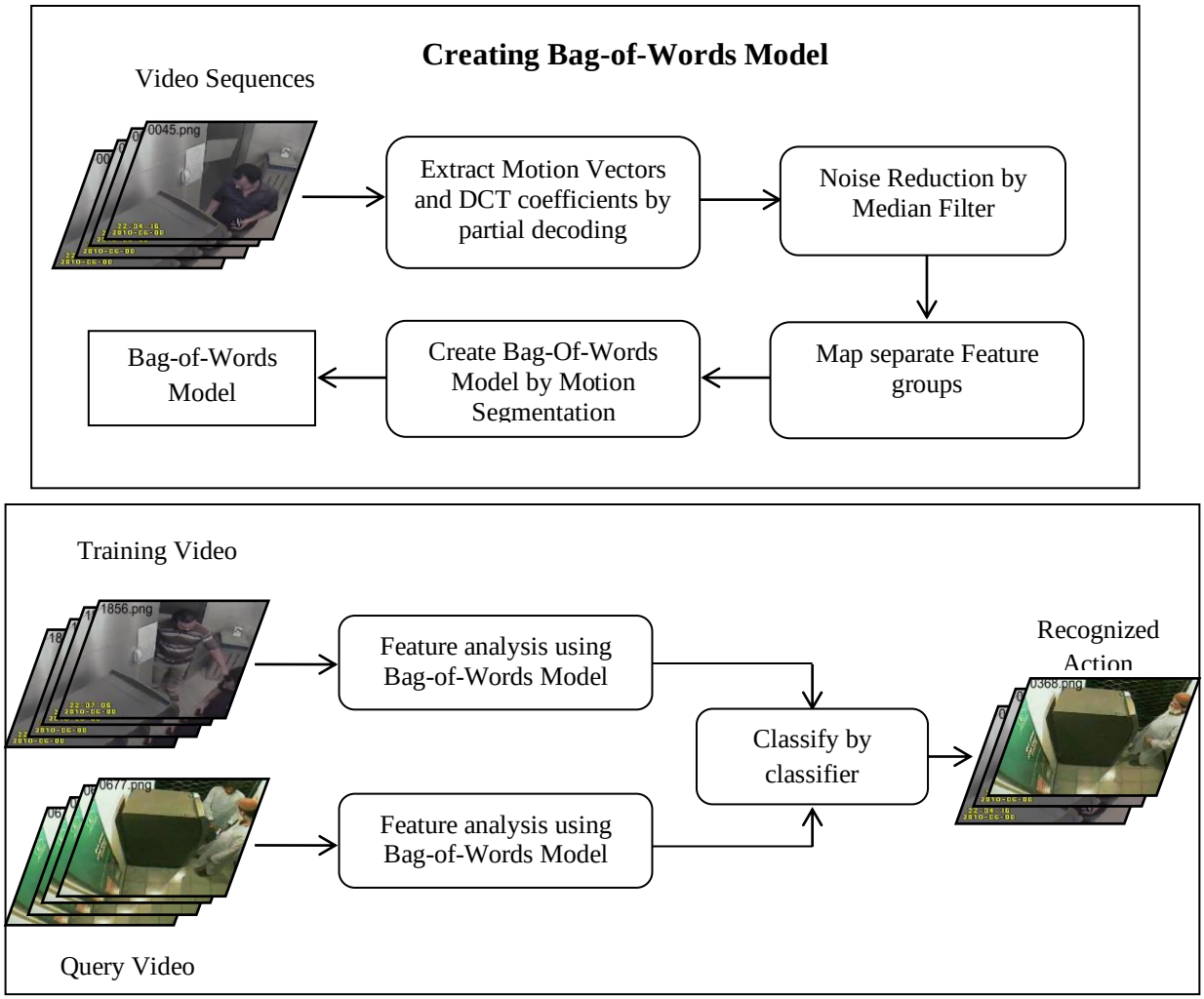


Figure 1: Functional Architecture of the Proposed System

Figure 1 show the proposed system that use compressed domain features such as Motion Vectors (MV) and DCT coefficients. Motion fields are accessed at the time of video decompression without additional cost for feature extraction. We can get sparse motion vector from motion estimation of video compression and are getting low storage space than pixel values and lower computational complexity. Then extracted sparse motion features are filtered by noise reduction to get dense features that are efficient for motion segmentation from avoiding inaccurate representation. Smoothed features are then mapped into specific separate groups according to motion direction. Finally, distinct groups are created as Bag-of-Words models that are histogram based descriptors used for video description. They are simple structures which encode a compact representation of the motion information.

4 MOTION FEATURE ANALYSIS

Motion features such as Motion vectors and DCT coefficients are analysis for efficient action recognition. Our system uses motion features that can be extracted by MPEG flow with slightly cost of partial decompression. Then extracted features are filtered and created bag-of-words model for activity recognition that aims to recognize the actions of one or more agents from a series of observations on the agents and the environmental conditions. It is a very important and challenging problem to track and understand the behaviour of agents through videos taken by various cameras.

4.1 Compressed Domain Features Extraction

Feature Extraction starts from an initial set of measured data and builds derived values intended to be informative and non-redundant. Our system extract motion features from video sequences that are composed of I frames, P frames and B frames. There are no motion data in I frames that own residue values. These values are DCT coefficients that the video encoders are not able to predict perfectly the changes in frame content. Motion Vectors (MV) are extracted from P and B frames. Motion vectors are quite prone to quantization error. Our system used to combine to motion vectors and DCT coefficients that are complementary to each other to get complete motion values.

Motion Vectors not only indicate the blocks under motion but also gives the information regarding magnitude and direction of the block with respect to the reference frame. DCT values are sum of sinusoids of varying magnitudes and frequencies. DCT values can get high residue data since the lower right values represent higher frequencies and much of the signal energy lies at low frequencies that appear in the upper left corner of the DCT. Motion vectors and DCT coefficients are shown in Figure 2. Figure 3 shows the orientation and magnitude of adjacent frames 0636, 0637 and 0638 at sampling rate of 25 frames per second.

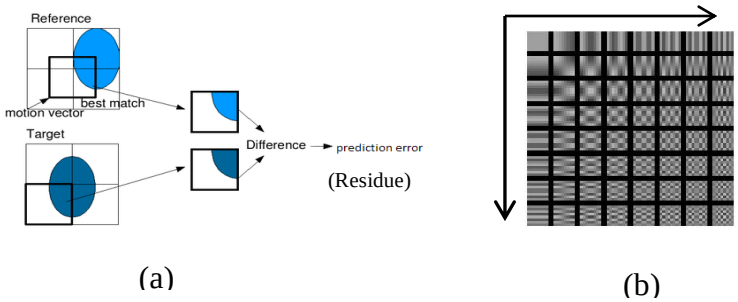


Figure 2: Compressed Domain Features: (a) Motion Vectors (b) Illustration of DCT with different residue values

(a) (b)

Figure 3: Example of Motion Vectors of Adjacent frames: (a)-(b)-(c)

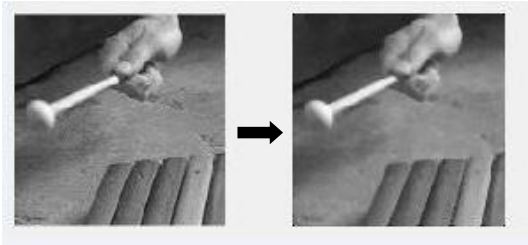
4.2 Motion Feature Filtering

Motion feature filtering in compression domain is very important for different noise reduction. Motion data that do not represent real values are needed to filter to be more reliable in motion segmentation. Our system uses nonlinear spatial median filtering, efficient for impulse noise reduction that arises as light and dark on the image. It preserves edges while removing noise as different mask shape such linear, square, circular, cross, and etc.

Median filter is a sliding-window spatial filter so it replaces the center value (i,j) in the window with the median of all the pixel values in the window. We present 2D Median filtering method in Equation (1).

$$g(x,y) = \text{med}\{f(x-i, y-j), \quad i,j \in W\} \qquad (1)$$

Where $f(x, y)$ is the original image and $g(x, y)$ is the output image, W is the two-dimensional mask: the mask size is n × n matrix. Figure 4 shows the comparison of original and filtered spatial frames.



(a) (b)

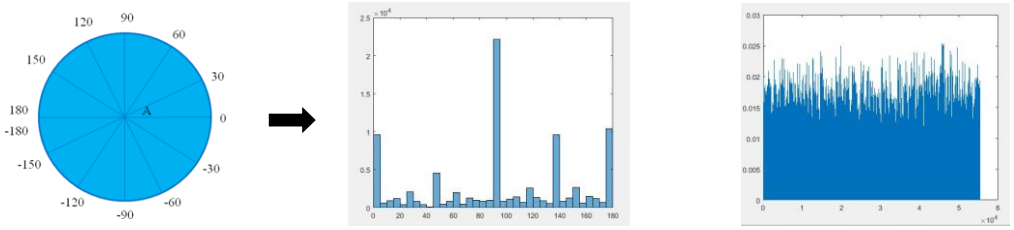
Figure 4: Example of Smoothing frames: (a) Original Frame (b) Median Filtered Frame

4.3 Mapping Filtered Compressed Features

We create separate histogram using the magnitude and orientation of filtered compressed features. The orientation values are counted in the range of 0-180 degree unsigned and inverse 0-180 degree from the center. There are 12 orientation bins with each of 30 degree as shown in figure 5 (a). The displacements on separate bins become magnitude to map distinct histograms as shown in figure 5 (b). The magnitude and orientation are calculated according to equation (2) and (3).

Magnitude: $$S = \sqrt{S_x^2 + S_y^2} \qquad (2)$$

Orientation: $$\theta = \arctan\left(\frac{S_y}{S_x}\right) \qquad (3)$$



(a) (b) (c)

Figure 5: Mapping Filtered Compressed Features: (a) 12 separate bins with each of 30 degree (b) Sample histogram of separate bins (c) Histogram of test frame with different virtual words

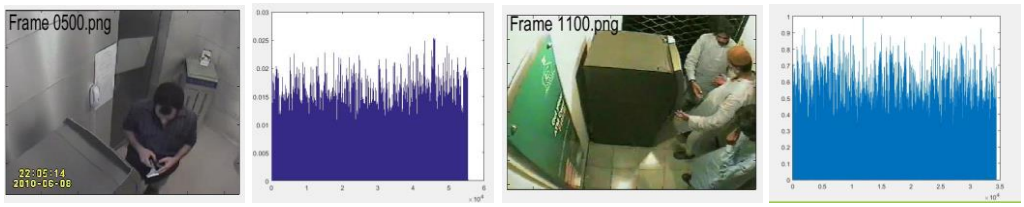
The extracted features are returned in a 1-by-*N* vector, where *N* is the histogram of oriented gradient (HOG) feature length. The histogram length corresponds to the number of visual words that becomes a feature. The returned features encode local shape information from regions within an image. Figure 5 (c) shows the histogram of test frame with different virtual words. We can use this information for many tasks including classification, detection, and tracking.

4.4 Bag-of-Words Model

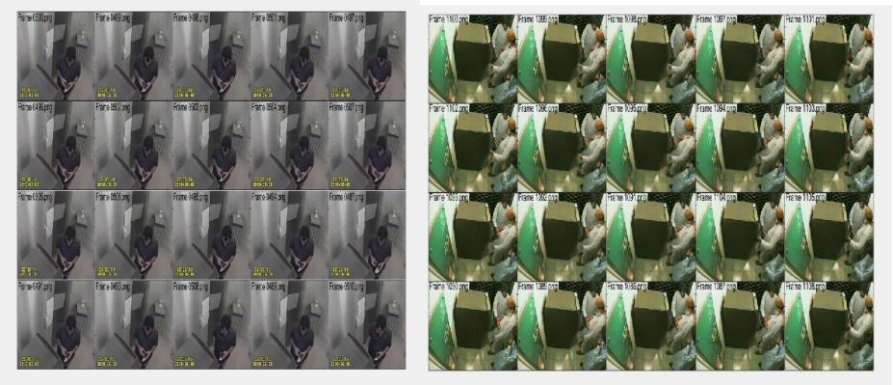
Bag-of-words model can generate a histogram of visual word occurrences that represent an image. That model can be applied to image classification by treating images in computer vision. The bag of features object defines the features, or visual words, by using the clustering algorithm. The resulting clusters are compact and separated by similar characteristics. Our system creates the bag of words model with different motion segments using k means clustering to recognize the human action more accurately.

5 SIMULATION EXPERIMENTS

In our system, compressed domain features are filtered and segmented into homogeneous subsets that can be more interpreted as a distinct feature. Then we create bag of words model by different features. We create the dataset from 3 different videos with the frame rate of 25 frames per second. We spite the overall data set with the ratio of 70:30 for training and testing by k fold cross validation with the value of k is 3. The first image dataset consists of 2,401 frames with 334,080 features and 1K clusters. The second image dataset consists of 3,930 frames with 547,056 training features and 1K clusters. The most match results can get during top 20 frames. Simulation results are change according to search range and number of features include in training sets.



(a) (b) (d) (e)



(c) (f)

Figure 6: Experiments: (a),(d) Query Image (b),(e) Segmented Features (c),(f) Query matched results

6 CONCUSION AND FUTURE WORK

We tend to detect all occurrences of query video sequence in a test video, and thereby recognizing an action as taking place at some specific time and invariant. In this paper, we used compressed domain parameters that can reduces the amount of effort involved in full decompression and keeps the storage cost low. Then motion filtering method is applied to reduce noisy motion values that can effect in motion segmentation .Then Bag-of-Words model is created using the distinct motion segments for specific feature description. Our system provided significant improvement in event analysis for surveillance application with small additional computational costs for efficient human action recognition in ATM installation.

Our system tends to use for security not only in ATM camp but also in different application areas such as traffic monitoring, crowded regions and royal places. We are trying to get the most effective motion segmentation technique for a future work.

REFERENCES

[1] M.Tom and R.V.Babu, "**Compressed Domain Human Action Recognition in H.264/AVC Video Streams**", Multimedia Tools and Applications, Springer, Volume 74, Issue 21, pp 9323-9338, Nov-2015.

[2] J.Miao, X.Xu, R.Mathew and H.Huang, "**Residue Boundary Histograms For Action Recognition In The Compression Domain**", IEEE International Conference on Image Processing, 27-30 Sept. 2015.

[3] R.V.Babu and K.R.Ramakrishnan, "**Recognition of Human Action using Motion History Information Extracted from The Compressed Video**", Image and Vision Computing , 597-607 , Nov-2003.

[4]C.Yeo, P.Ahammad, K.Ramchandran and S.S.Sastry , "**Compressed Domain Real-time Action Recognition**", IEEE Workshop in Multimedia Signal Processing, 2006.

[5] K.Rapantzikos and M.Zervakis, "**Robust Optical Flow Estimation In MPEG Sequences**", IEEE International Conference on Acoustics, Speech and Signal , Volume 2, 2005.

[6] V.Kantorov and I.Laptev, **"Efficient Feature Extraction, Encoding and Classification for Action Recognition**", Computer Vision and Pattern Recognition (CVPR), IEEE, 2014.

[7] J.Uijlings, I.C.Duta, E.Sangineto and Nicu Sebe, "**Video Classification with Densely Extracted HOG/ HOF/ MBH Features: An Evaluation of the Accuracy/ Computational Efficiency Trade-off",** International Journal of Multimedia Information**,** Volume 4, Issue 1, pp 33-44, March 2015.

[8]D.K.Vishwakarma, A.Dhiman, R.Maheshwari and R.Kapoor, **"Human Motion Analysis by Fusion of Silhouette Orienteation and Shape Features" ,** 3rd international Conference on Recent Trends in Computing, (ICRTC), 2015.

[9]J.Pan ,S.Li and Y.Q.Zhang , "**Automatic Extraction of Moving Objects Using Multiple Features And Multiple Frames**", IEEE Institute of Electrical and Electronics Engineers, April-2000.

[10] V.Tripathi, D.Gangodkar, V.Latta and A.Mittal , "**Robust Abnormal Event Recognition via Motion and Shape Analysis at ATM Installations**", Journal of Electrical and Computer Engineering, Volume 2015, Article ID 502737,10 pages, 2015.

[11] A.Barjatya, "**Block Matching Algorithms For Motion Estimation**", DIP 6620 Spring 2004 Final Project Paper, IEEE