Compressed domain Action Recognition for Healthcare and Assisted Living

Ali Abdari
Department of Engineering
Faculty of Electrical and Computer Engineering
Kharazmi University
Tehran, Iran
Email: std_ali.abdari@khu.ac.ir

Pouria Amirjan
Department of Engineering
Faculty of Electrical and Computer Engineering
Kharazmi University
Tehran, Iran
Email: std_pamirjan@khu.ac.ir

Azadeh Mansouri
Department of Engineering
Faculty of Electrical and Computer Engineering
Kharazmi University
Tehran, Iran
Email: a_mansouri@khu.ac.ir

Abstract— **Traditional action recognition methods are time-consuming and need a high-performance hardware for required calculations. Nowadays in many popular applications, compressed videos are available. We proposed a method which uses available information in compressed domain and improves the performance of extracting features from neural network by using residuals of frames instead of decoded and reconstructed frames. This work proposes a fast and efficient recognition of activities. The proposed approach reduces the computational cost of action recognition since the compressed video information is explored. Low complexity of the proposed method makes it proper specially for healthcare and assisted living purposes. The experimental results clearly in daily living dataset illustrate that the proposed low computational compressed domain approach provides an acceptable performance in terms of recognition accuracy.**

.

Keywords-action recognition; compressed domain; ADL ;real time applications.

I. INTRODUCTION

As the aging population increases in modern communities, the need for remotely protecting elderly and patient at the home or other places has recently increased. Generally, the aim of the healthcare and assisted living is improving the quality of life and providing healthy living of older or impaired people by using information technologies and machine vision. Technology can support people affected by various physical or mental disabilities such as chronic disease or elders who have Alzheimer.

In the last decade artificial intelligence has emerged as a powerful tool and can be useful in many aspects of life, such as healthcare and assisted living. Smart technologies can be utilized as a means to improve the quality of care and wellbeing of dependent people [1].

For some of applications such as public security usage, video-based monitoring technology is highly developed. This attention is now focused on the use of this technology for healthcare and assisted living. A range of applications, such as human activity recognition, gait analysis and fall detection are some of these products. Recognition of Activities of Daily Living (ADL) has drawn significant research's attention in the last decade. Monitoring these activities can provide valuable information for applications such as assisted living, remote healthcare, and lifestyle [2].

Generally, using wearable sensors and online processing of recorded videos are two common approaches for monitoring people. Wearing sensors can cause an uncomfortable situation for elders or patients. Actually, deploying new low cost monitoring systems for old people and patients makes their living places smart environments which is inevitable these days since the aging population increases [3]. Many places like hospitals, equipped houses or places for caring disabled people have been surrounded with installed cameras. The recorded videos can be utilized to reduce the cost of preparing wearable sensors. In this case, there is no need for patients or monitored people to carry extra components.

As it is mentioned, one of the unobtrusive and efficient methods for activity recognition and monitoring is video based approaches. Providing real time method which require only some information derived from recorded videos is of primary importance. Extracting appropriate features like optical flow for video based action recognition and monitoring is a too time-consuming task and it is not practical to use this type of features for real-time applications. On the other hand, most cameras utilize a kind of compression during recording videos; since the required capacity for storing raw video is severely high. Moreover, compressed domain videos can be better employed for transferring data through networks. More recently, since most of the monitoring systems assumed to be analyzed over digital network, the compressed domain action recognition and monitoring has been interest in more applicable designs. Generally, extracting features using compressed domain information can result in faster and more applicable action detection, recognition and monitoring systems.

II. RELATED WORK

As we mentioned before, using wearable sensors and video-based monitoring technology are two common approaches for monitoring daily living activities.

Usually, sensors contain accelerometers or gyroscopes. As an example for monitoring fall detection event, there are many types of sensors including measuring the vibration of the floor [4], detecting a fall event by using the pressure mats [5] or impulse-radar sensors [6].

Practically, there has been considerable researches devoted to action recognition in the last years. This field has many applications in several domains such as intelligent video surveillance, video retrieval, video content analysis and activities of daily living (ADLs). Recognition of activities of daily living refers to detect the activities that commonly happen in daily life. Classifying this type of activities are of particular interest in many applications, such as health monitoring, smart home environments and video surveillance systems.

Motion and appearance information are two important features for action recognition [7]. This information usually is extracted using row videos. In the following firstly the traditional methods are explained briefly and then the compressed domain approaches are described.

A. Traditional approaches

In common approaches the handcrafted motion features are usually extracted using optical flows. Calculating optical flows is very time-consuming process which restrict the usage of this features for real-time applications. For example, in [8] authors use the motion boundary of optical flows for a dense sampling of interest points. Interest points are tracked over time by an enhanced Kanade Lucas Tomasi (KLT) tracker [9] and accumulated in a three-dimensional trajectory structure. In this case, multiscale descriptors are employed for representation. In some video stream, the large variation between frames may lead to misclassifications. In [10] the authors propose a multiple subsequence combination (MSC) method that divides the video into several consecutive subsequences. Authors in [11] proposed a method to employ salient proto-objects for unsupervised discovery of object and object-part candidates using HOF[1] to achieve the best result. In [12] authors present a hybrid approach for long-term human activity recognition in a more precise manner compared to unsupervised approaches. These methods employed optical flow and therefore cannot be utilized in real time applications.

Considering the high redundancy of the video frames; in [13] authors utilize the difference of frames for training the deep CNN. In this work, the energy consumption is reduced without much degradation in accuracy since the motion can be estimated roughly through difference of frames.

On the other hand, compressed videos as available information can be directly utilized for extracting the compressed domain features. Moreover, carefully selection available information can provide acceptable performance for real-time applications. In the next sub-section, a brief description about compressed domain information is given at first and then some of existing compressed domain methods is described.

B. Compressed domain information

Nowadays, the necessity for video transmission has led to storing and transmitting the video files in compressed form. All of the video compression algorithms take the advantage of the fact that the successive frames are highly redundant. As a result, most of them split a video into I-frames (intra frames), P-frames (predictive frames) and B-frames (bi-directional frames). I-frames are considered as an image and compressed like a still image. For P and B frames, considering the encoded reference frame(s), just the changes are encoded and transmitted. As it is illustrated in Figure 1 there are two important components as motion vectors (MV) and quantized coefficients. MV's shows the movements of block of pixels from the source frame to the target frame. The difference between the original block and the best match block is referred as residual difference. Discrete Cosine Transform (DCT) is applied on the residuals in order to obtain quantized coefficients.

Motion can be used for action recognition like optical flow. Motion vectors is block based and it is not as accurate as optical flow, but it can be useful for real time applications since by using compressed video, this information is freely accessible. Apart from that, residuals can be achieved by applying inverse DCT to available quantized coefficients. Actually, these values illustrate the difference of consequent frames and have much important information for action detection.

C. Compressed domain approaches

In [2] Faster motion estimation methods are employed instead of costly dense optical flow based motion estimation, by using the motion vectors which are accessible directly from the compressed videos. This type of approaches reduces the computational complexity, with minimal loss in terms of recognition accuracy.

Motion Vectors (MV) are calculated for each macro-block of size M×M and therefor precisely evaluating the real motions through MVs cannot be attainable. On the other hand, optical flow is pixel based and it is more accurate than motion vectors, but extracting optical flow is not appropriate for real time applications, however optical flows can be used in offline cases. In some approaches like [14] the authors train their model with optical flow and fine tune the weights of network by motion vectors in training phase. Then for test step, they just use motion vectors as input for model. By using this approach, the accuracy of method will be close to optical flow based method and the speed will increase as well.

Discrete Cosine Transformation (DCT) includes important information like high frequency and low frequency

[1] Histogram Of Flow

range. In [15] authors obtain a feature vector from horizontal and vertical and DC features from available DCT blocks in residues in compressed domain.

Residues in a compressed video are most likely to the differences of frames in a raw video since residues contains essential information about spatial and temporal domain. This information can be so useful for action detection and recognition. We proposed a new method to use available residuals in a compressed video. First, we decode video partially and calculate the inverse of DCT coefficients of every encoded block, then the achieved results are concatenated to reconstruct frames of residuals. After that, we use the obtained video as an input for convolutional neural network (CNN) for extracting features. Finally; a support vector machine (SVM) classifier is employed to detect the class of occurred action.

The structure of the paper is presented as follows: The proposed method is expressed in section III. Finally, in section IV obtained results on KIT dataset is illustrated and the section V is dedicated to the conclusion.

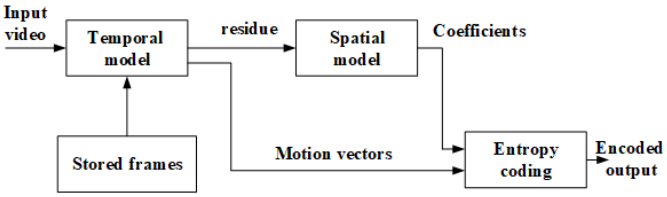

Figure 1. Video compression model

III. PROPOSED METHOD

Compressed videos already carry motion information such as MV's and residuals which are very useful for action recognition. While compression process of a video, frames will divide to some blocks, and for each block a best match will be found by searching in a reference frame. Then the difference between every block and the related best match block will be calculated. The quantization of discrete cosine transform of differences will be encoded and stored in compressed video format. So for every block a residual and 2-D vector which represents the motion vectors of the block are generated and transmitted.

The size of blocks depends on the type of compression standards; in some compression method the size of the blocks is consistent and in the recent standard it is feasible to slice the frames by different sizes of the blocks; regions with fast motions and complex structures will be encoded by fewer sizes of blocks and regions with simple structures will be encoded by larger sizes of blocks. For example, in MPEG2 standard the size of blocks is considered as 16×16.

In the proposed method we assume that a compressed video file is available. Then the compressed video is partially decoded to obtain residues. In this case, we should calculate the inverse discrete cosine transform to achieve residues in spatial domain. Now, we use obtained residue as an input of a pre-trained neural network.

A pre-trained neural network as a feature extraction tool is considered as a parts of the proposed method. In this paper, ImageNet VGG-f is considered as the pre-trained model which is trained using MatConvNet[2] on imagenet dataset Figure 2. Shows the architecture of this model that contains 21 layers.

The input of the model is the residual which is obtained by partially decoding the compressed domain videos. These inputs are resized into 224 × 224 for network training. Moreover, the normalization step is considered based on a calculated average image that specified in the pre-trained network. Outputs of the 18th layer of size 4096, is considered as the extracted feature. Every input produces a 4096×1 feature vector. As a result, an input video of length M will generate a matrix of size 4096×M.

For classification step we used SVM classifier with X^2 kernel for action classification. The size of Input in each classifier is unique, but the output of the feature extraction step has size 4096×M. In order to reshape the matrix to a unique size, we use Pooled Time Series approach [16]. In this method, a time series pooled by a pooling operator such as maximum pooling. The output of this operator is a vector of size 4096×1 for each video. In addition, we conducted a technique for data augmentation and action segmentation. We split a video into several batches and apply max pooling on it. The action of each batch recognized and the class of the input video obtains from voting among the recognized action of batches.

For example, if we consider batch size as 8, the main video split into eight segments. Pooling method applied on each batch and produce eight decision. For classify the action of the input video, we vote among this decisions.

Using the proposed CNN based approach important and vital features from the residues frame can be extracted.

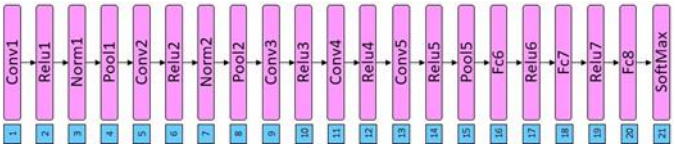

Figure 2. Layers of VGG Neural Network

- Input a compressed video
- Partially decode the video applying IDCT in order to obtain residuals
- Concatenate reconstructed residual frames to build a video
- Split the reconstructed video to desired batch size
- Extracting features for every batch using the proposed CNN
- Use classifier to detect the action

[2] As explained in "http://www.vlfeat.org/matconvnet/"

Figure 3. The summary of proposed method

IV. EXPERIMENTAL RESULTS

For analyzing the performance of our method we test it on KIT dataset [17]. As we see in Figure 4, KIT is a third person dataset, it is provided in the kitchen and has 10 classes includes: clear table, drink coffee, cut vegetables, empty dishwasher, peel vegetables, eat pizza, set the table, eat soup, sweep the floor and wipe the table. There are about 17 video in every class and the size of videos is 480×640.

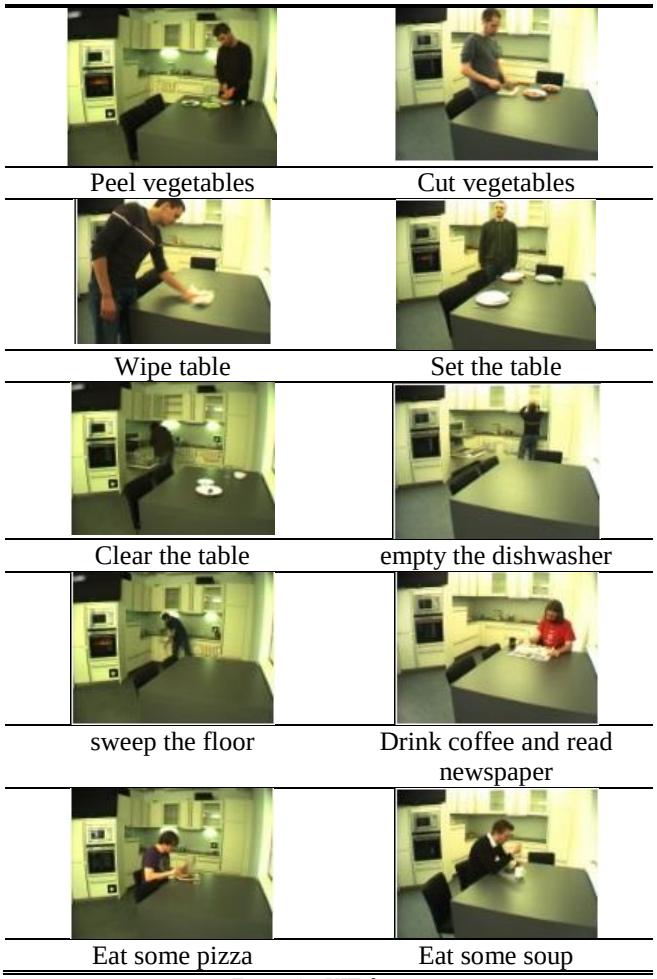

Figure 4. KIT dataset

We assume that the compressed video in (MPEG-2) format is available. Then, we decoded a video partially, and concatenated the residues to obtain a residual frame. For enhancing the performance and speed the video sequences are temporally down sampled by a factor of five. After that we divide the video into the number of desired batches, then we set the frames of every batch as an input for mentioned CNN and extract features from the 18'th layer.

When the features for a batch obtained, we used pooled time series approach order to reshape the matrix to a unique size. For training the SVM classifier, we use half of the videos per class for training and the rest for the test. The final accuracy of each classes are illustrated in Table 1 which obtained using the mean values of 100 times train and test by shuffled data. We compared the performance of the proposed method against the three approaches [8], [10] and [11].

In [11] the authors proposed a method that employ salient proto-objects for unsupervised discovery of object and object-part candidates and combined it by HOF to achieve the best result. They utilize a linear multi-class SVM and train a multinomial logit model on the training data via cross-validation. Also, all features are standardized to zero-mean and unit-variance, since this feature scaling method proved to yield robust results. The method proposed in [10] recognizes the human action in a video by combining the classifications of its multiple subsequences. The input is divided into a number of consecutive and non-overlapping subsequences of a given length and then classify the action performed in each subsequence. In the end, the authors combine subsequence classifications to obtain the final label of the video. Their descriptors are based on gradients and optical flow. For a better decision of whole video class, they used MES framework [18], [19], [20]. In [8] the authors use Motion boundary histogram and KLT tracker for detecting meaningful regions in frames. Also they used SVM classifier with a Chi-Square kernel for detection purpose. It seems that they split data to 66% for train and the others for the test phase. In Table 1 Max_1 to Max_8 is corresponds to size of batch from 1 to 8 for splitting the videos.

Table 1. Experimental results of the proposed method in comarison with the exsiting methods

Actions	Max_1	Max_4	Max_8	HOGHOF [7]	[9]	proto-object + HOF [10]
Clear table	**98.2%**	92.3%	94.2%	**100.0%**	**100.0 %**	94.4%
Coffee	**93.2%**	92.2%	91.4%	**100.0%**	**100.0 %**	91.2%
cut vegetables	59.5%	68.3%	**70.3%**	**83.3%**	**71.0%**	65.2%
empty dishwasher	93.0%	98.0%	97.3%	**100.0%**	**100.0 %**	**98.8%**
Peel	63.0%	**75.5%**	**79.8%**	**88.2%**	57.0%	72.6%
eat pizza	61.4%	65.3%	69.3%	**78.6%**	**86.0%**	**88.3%**
set table	90.4%	91.8%	**94.6%**	92.8%	**100.0 %**	**98.6%**
Soup	67.1%	80.1%	**83.9%**	**92.8%**	71.0%	**88.2%**
Sweep	99.8%	100.0%	**99.6%**	**100.0%**	**100.0 %**	88.7%
Wipe	89.0%	97.8%	**97.0%**	**100.0%**	**100.0 %**	93.4%

The per class accuracies are illustrated in table 1. As it is clearly shown, the top three results per class is depicted in bold face. In some classes such as Peel and cut or pizza and soup all of the methods achieve less accuracy. This results are obtained since the actions are similar. Although the complexity of the proposed method is less than the high complexity optical flow based approaches, the results for these challenging actions are acceptable specially for Max-8 implementation. There is always a tradeoff between accuracy and computational cost. The fact is that, the accuracy of any method may be limited to speed. Exploring all the results depicted that although the proposed method provides slightly lower accuracy of some classes in comparison with other computational expensive methods; utilizing available

information in the proposed compressed domain method provides the low computational complexity approach which makes the algorithm appropriate for real-time applications such as healthcare and assisted living.

V. CONCLUSION

In the proposed compressed domain method, the residuals are obtained using partially decoded quantized coefficients as the available elements of the compressed domain video. Residuals contain very useful data such as temporal and also spatial information. The residuals then are fed into CNN in order to extract appropriate features for action detection and recognition. In this case, the time of extracting features is reduced since we do not need the whole process of decoding Practically, residues can be obtained by partially decoding the videos. The low complexity of the proposed method makes it suitable for real time application such as healthcare purposes.

REFERENCES

[1] M. Kepski and B. J. I. C. V. Kwolek, "Event-driven system for fall detection using body-worn accelerometer and depth sensor," vol. 12, pp. 48-58, 2017.

[2] S. Poularakis, K. Avgerinakis, A. Briassouli, and I. J. S. P. I. C. Kompatsiaris, "Efficient motion estimation methods for fast recognition of activities of daily living," vol. 53, pp. 1-12, 2017.

[3] Y. Hbali, S. Hbali, L. Ballihi, and M. Sadgal. (2018, Skeleton-based human activity recognition for elderly monitoring systems. *IET Computer Vision 12(1)*, 16-26. Available: http://digital-library.theiet.org/content/journals/10.1049/iet-cvi.2017.0062

[4] D. Litvak, Y. Zigel, and I. Gannot, "Fall detection of elderly through floor vibrations and sound," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 4632-4635.

[5] Z. Zhang, U. Kapoor, M. Narayanan, N. H. Lovell, and S. J. Redmond, "Design of an unobtrusive wireless sensor network for nighttime falls detection," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 5275-5278.

[6] R. Z. Morawski, Y. Yashchyshyn, M. Piórek, F. F. Jacobsen, K. Ovsthus, and W. J. T. R. T. N. Winiecki, "Monitoring of human movements by means of impulse-radar sensors," vol. 6, pp. 598-602, 2015.

[7] L. Wang, Y. Qiao, and X. J. T. A. R. C. Tang, "Action recognition and detection by combining motion and appearance features," vol. 1, p. 2, 2014.

[8] K. Avgerinakis, A. Briassouli, I. J. J. o. A. I. Kompatsiaris, and S. Environments, "Activities of daily living recognition using optimal trajectories from motion boundaries," vol. 7, pp. 817-834, 2015.

[9] H. Wang, A. Kläser, C. Schmid, and C.-L. J. I. j. o. c. v. Liu, "Dense trajectories and motion boundary descriptors for action recognition," vol. 103, pp. 60-79, 2013.

[10] L. Onofri, P. Soda, and G. J. I. C. v. Iannello, "Multiple subsequence combination in human action recognition," vol. 8, pp. 26-34, 2014.

[11] L. Rybok, B. Schauerte, Z. Al-Halah, and R. Stiefelhagen, "“Important stuff, everywhere!” Activity recognition with salient proto-objects as context," in *2014 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014, pp. 646-651.

[12] F. Negin, M. Koperski, C. F. Crispim, F. Bremond, S. Coşar, and K. Avgerinakis, "A hybrid framework for online recognition of activities of daily living in real-world settings," in *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*, 2016, pp. 37-43.

[13] B. Han and K. J. I. T. o. M.-S. C. S. Roy, "DeltaFrame-BP: An Algorithm Using Frame Difference for Deep Convolutional Neural Networks Training and Inference on Video Data," 2018.

[14] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. J. I. T. o. I. P. Wang, "Real-time action recognition with deeply transferred motion vector CNNs," vol. 27, pp. 2326-2339, 2018.

[15] J. Miao, X. Xu, R. Mathew, and H. Huang, "Residue boundary histograms for action recognition in the compressed domain," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 2825-2829.

[16] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled motion features for first-person videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 896-904.

[17] L. Rybok, S. Friedberger, U. D. Hanebeck, and R. Stiefelhagen, "The kit robo-kitchen data set for the evaluation of view-based activity recognition systems," in *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, 2011, pp. 128-133.

[18] T. K. Ho, J. J. Hull, S. N. J. I. t. o. p. a. Srihari, and m. intelligence, "Decision combination in multiple classifier systems," vol. 16, pp. 66-75, 1994.

[19] J. Kittler, M. Hatef, R. P. Duin, J. J. I. t. o. p. a. Matas, and m. intelligence, "On combining classifiers," vol. 20, pp. 226-239, 1998.

[20] L. I. Kuncheva, J. C. Bezdek, and R. P. J. P. r. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," vol. 34, pp. 299-314, 2001.