Human Activity Recongnition-A survey

Aral Sarrafi
Department of Mechanical Engineering
University of Massachusetts Lowell
Lowell, Massachusetts, 01854
aral_sarrafi@student.uml.edu

***Abstract*: Activity recognition can be defined as labeling videos or sequence of images. Efficient solutions to this problem can have several applications in security, public crowd control, Robotics, surveillance, content-based video indexing, and interactive environments. Activity recognition is a challenging task because of wide range of variabilities associated with spatial-temporal patterns. This survey paper, will present an overview of the several methodologies and techniques in activity recognition, including the classical approaches and the deep learning method which is the new trend in the computer vision community for activity recognition.**

Keywords: Human Activity Recognition, Convolutional Neural Network, Deep Learning

I. Introduction

Activity recognition is a subset of the general video classification problem. In activity recognition context the videos that are being studied contain human motion and activity [1]. Action is defined as a sequence of movements generated by a human agent during the performance of a task. Action recognition can be described as determining the action label that best describes an action being performed in a video scene by a different agent under different viewpoints and potentially different manner, speed or style. This challenging problem has several real-time and off-line application. For example, automatic surveillance and security in shopping malls and airport are the most highlighted real-time applications of activity recognition. The video game companies have also utilized some of the technologies in activity recognition to create interactive environments to provide a better gaming experience for the user [1, 2].

Human activity recognition is closely linked to other topics in computer vision such as motion estimation, tracking, image classification and human motion analysis. The major components of most of the action recognition can be considered as feature extraction, learning and classification and action segmentation [3].

Feature extraction is the most important and the very first step in all different classification problems. Ideal Features are parameters associate the with the data set that are sensitive to a specific concept and ideally invariant to unrelated concepts [2]. In the context of the activity recognition the features are considered as extracting posture and motion cues from the video or sequence of images that are sensitive to human action that are being performed in the video. Ideally, the features should be robust to different variabilities such as partial occlusion, background clutter, shadows and different illumination conditions [3]. After obtaining reliable features, the learning and classification step comes into play. The learning can be represented as a statistical model from the extracted features, and then the trained model can be utilized to classify the new feature observations. A major challenge during the model selection and learning procedures is the large variabilities associated with several actions that are being performed with different human agents or under different conditions.

Action segmentation is also another important problem to address in the context of the activity recognition, in order to cut the streams of the video into single action instance. The action segmentation can provide a reliable data set the learning procedure as well.

In recent years due to availability of large amount of digital data and high computational power in advanced GPU processors, deep learning approaches have been gaining more attention in all different areas to solve all different types of classification problems including the activity recognition. Using an End-to-End deep learning approach enables the algorithm to skip the feature extraction stage of the activity recognition. In other words, the deep neural network will be able to learn the suitable features for activity recognition from the large amount of data set. Therefore using the hand-crafted features will not be necessary while working with deep learning.

This survey paper will be reviewing the activity recognition problem from the both classical video classification and deep learning perspective. First some of the classical approaches will be introduces.

II. Classical Approached

A. Temporal templates

In the temporal template approach the objective to map the videos as 3D signals into 2D images, while preserving the temporal information , some of the successful methods in the temporal template approach are motion energy images (EMI) and motion history images (MHI) [4]. A brief overview on the EMI and MHI will be presented later for clarification.

It should be highlighted that the motion estimation methods plays a major role in the classical approaches. For example for computing both the motion energy images and motion history images determining the pixels that have moved at each frame is essential.

Let's, take closer look at the formulation of the motion energy images. Let's assume that $D(x,y,t)$ is the binary image sequence indicating regions of motion. In other words $D(x,y,t)$ will be 1 at the pixels that have move in that specific frame, and it will be zero elsewhere. The motion energy images can be formulated as below.

$$E_\tau = \bigcup_{i=0}^{\tau-1} D(x,y,t-i) \tag{1}$$

Figure below shows an example of motion energy image formation for an example video of a person while sitting on a chair.

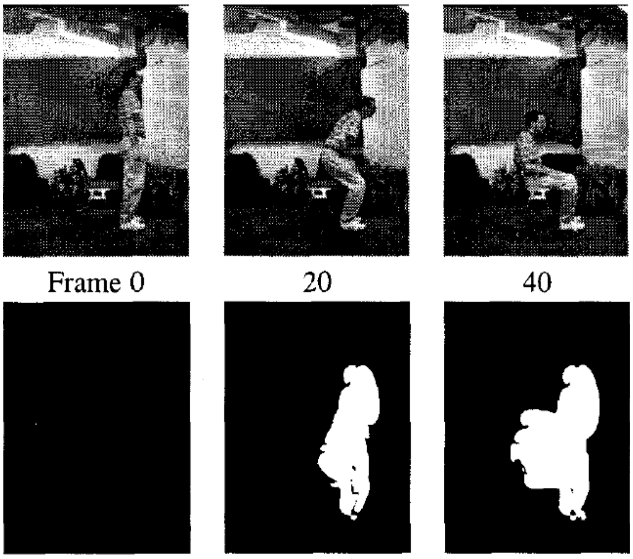

Figure 1.Example of motion energy image (MEI)[4]

The motion energy image can be a useful hand-crafted feature for activity recognition. However, the main draw-back associated with the motion history images is that it loses all the temporal information. The EMIs only takes into account the pixels that have moved without considering then moment in time in which the pixel has mover. Therefore, other features that are able to take into account the temporal information can be more useful for activity recognition.

Motion history images are another way of mapping 3D videos on to 2D feature maps. The motion history images can be formulated as below:

$$H_{\tau}(x,y,t)=\begin{cases}\tau & if \quad D(x,y,t)=1\\ max(0,H_{\tau}(x,y,t-1)-1) & otherwise\end{cases} \quad (2)$$

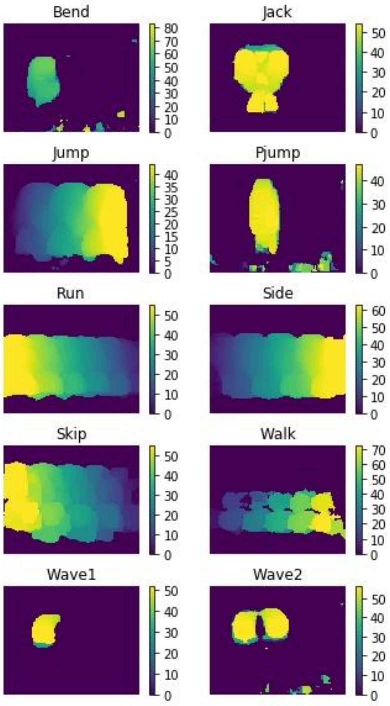

Figure 2. Examples of motion history image for different activities of the WIZEMANN data set

The advantage of motion history images over motion energy images is that the motion history images will preserve the temporal information of pixels that have moved. Figure 2 shows the motion history image for the 10 different actions of the WIEZMANN activity recognition data set.

As it is clear each of the actions will be mapped to its corresponding motion history image, and the activity recognition problem will be reduced to a 2D images classification which has been studied for years in the literature of computer vision.

The classical approaches for classification of the motion history images and activity recognition usually used the Hu moments to reduce the problem even further. Hu moments are scale and translation invariant features for images. Therefore, by computing the Hu moments for each of the motion history images associated with each activity all the videos can be represented by a feature vector, which consists of the Hu moments. These feature vectors can be classified with state of the art classification networks such as support vector machines (SVM).

III. Convolutional neural networks and deep learning approaches

Convolutional neural networks have already revolutionized different areas in the computer vision such as object detection and object recognition, and the same ideas can be generalized for video classification and action recognition as well. In the context of image classification, object recognition and object detection the 2D convolutional layers plays a major role. When it comes to video classification the 2D convolutional layers should be replace with 3D convolutional layers because of the additional temporal component that is associated with videos. Figure below shows an intuitive comparison between the 2D and 3D convolution operator.

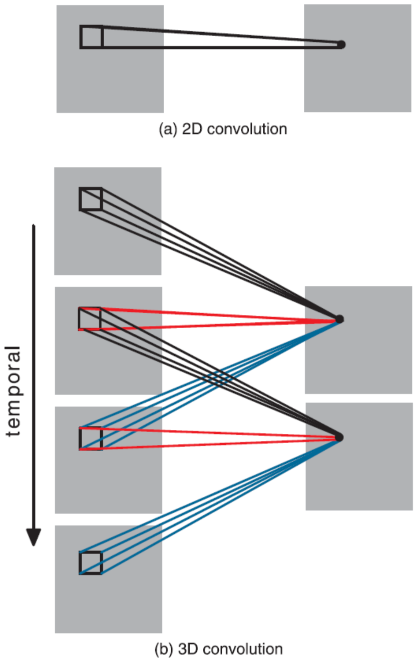

Figure 3. Comparison between the 2D and 3D convolution operator. a) 2D convolution b) 3D

A. *3D convolutional neural network*

One of the successful architectures of 3D convolutional deep neural networks were proposed by Ji *et.al* [5]. The proposed model was evaluated on the TREC Video Retrieval Evaluation (TRECVID) and KTH activity recognition data. This data set consist of surveillance video data recorded at London Gatwick Airport. The proposed activity recognition pipeline 3D convolutional network for surveillance event detection.

The proposed methodology by Ji *et.al* consist of a hardwired layer prior to the convolutional layers. The hardwire layer takes 7 frames of the video and constructs the gray scale frames, gradient in x direction, gradient in y direction, optical flow in x direction and optical flow in x direction. Therefore the seven frames will be hardwired to the feature maps which will be in total 33 feature maps. (3 layers 7 feature maps gray, gradient-x gradient-y and 2 layers with 6 feature maps including the optical flow-x and optical flow-y). After the hardwired layer the convolutional layers, subsampling layer and the fully connected layers will come into play as follows:

Convolution Layer 1: 23 filters, size 7*7*3
Subsampling Layer 1: size 2*2
Convolution Layer 2: 13 filters, size 7*6*3
Subsampling Layer 2: size 3*3
Convolutional Layer 3: 7*4*1
Fully Connected Layer: With 128 neurons

Because most of the human actions span a number of frames, some of the high-level motion information was also encoded to the model for regularization purposes. The performance of the algorithm was evaluated by several accuracy criteria including Precision, Recall, area under ROC curve (AUC) for multiple false-positive rate FPR.

B. *Convolutional Two-Steam Network Fusion*

Karen Simonyan and Andrew Zisserman introduced the original two stream deep neural network [6]. This networks consist of two separate streams for processing the video including the spatial stream ConvNet and the temporal stream ConvNet. The spatial stream ConvNet processes a single frame at a time from the video stream, and the temporal stream will process the estimated multi-frame optical flow Then the output of each of the streams will be fed to a class score fusion for the final label prediction. An example architecture of the two stream convolutional network is presented in figure 4.

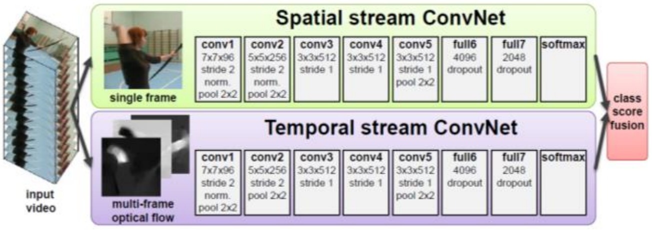

Figure 4. An example of a two stream network for video classification[6]

Feichtenhofer *et.al* proposed a new architecture for the two stream neural networks for activity recognition called the convolutional two-stream-network fusion for video action recognition [7]. They have reported that rather than fusing at the softmax layer, a spatial and temporal network can be fused at a convolutional layer without loss of performance while reducing the number of training parameters.

Moreover, this research has shown that it is better to fuse such networks spatially at the last convolutional layer rather than earlier, which will increase the accuracy. Figure 5 shows two examples of where a fusion layer can be placed efficiently based on this method.

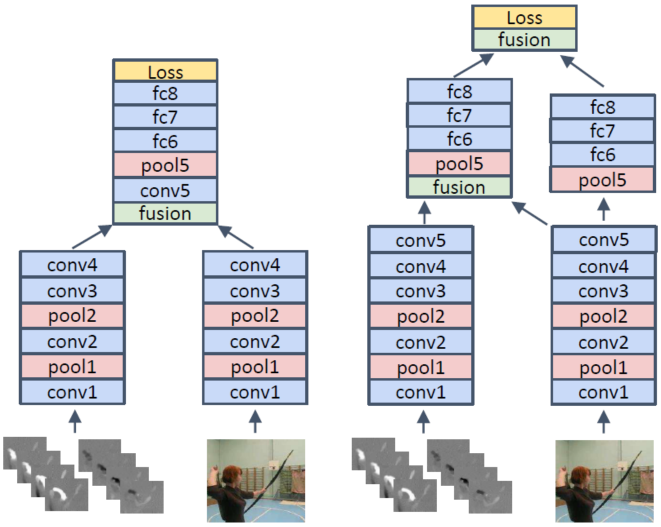

Figure 5. Potential locations for placing the fusion layer[7]

This spatiotemporal architecture for the two stream networks does not increase the number of parameters significantly compared to other methods while increasing the accuracy on the benchmark data sets. It has been shown that changing the location of fusion layer can be a good parameter to play with in order to achieve a better performance with two-stream networks.

Another successful approach in two stream networks was combining the object detection and object recognition approaches with temporal video processing for activity recognition [8]. As mentioned in the introduction the object detection and object recognition is a well-studied problem in the literature of computer vision and there are numerous well developed benchmark approaches to tackle this problem. The activities that are including presence of objects in the scene can provide more information to ease the predicting the correct labels for a video.

For example if there is a person in the scene with a bicycle these clues can be very helpful for the network to predict that the action that is being performed is ‘riding’. A schematic representation of this approach is present in Figure 6.

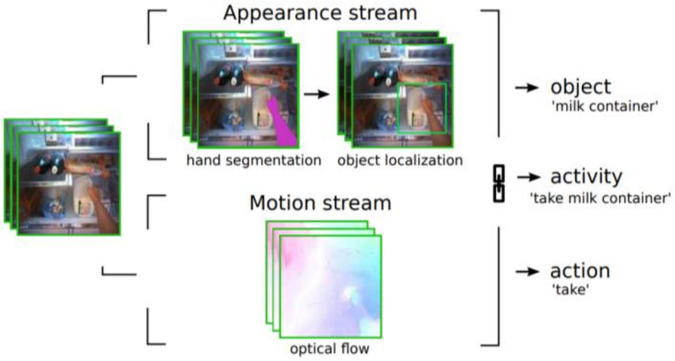

Figure 6. A schematic of the combining the object detection and recognition into the activity recognition network[8]

C. *Recurrent networks for visual recognition and description*

Generating descriptions for video is a much more challenging task in the context of the activity recognition. The main difference is that in activity recognition only a simple verb will be predicted for the observed video. On the other hand for discerption generation a complete sentence should be predicted by the network, which describes the video most accurately. This task can be perceived as an overlapping area in between the natural language processing and computer vision.

Generating the video description cannot be performed by only using convolutional neural networks, and sequence-learning layers should be also added to the final model. Figure 7 shows an example of hybrid network architectures for video description. As it is clear this model is a combination of convolutional layers and sequence learning layers [9].

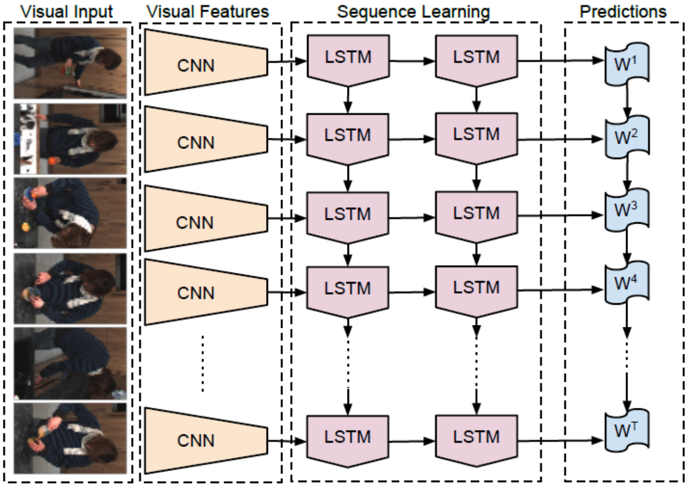

Figure 7. An example of a long-term recurrent convolutional network for video description generation[9]

It has been shown that the Recurrent networks for visual recognition and description with a deep architecture in both the spatial and temporal domain has the flexibility to be applied to several vision tasks involving sequential inputs and outputs. The results with are showing that by learning the sequential dynamics with a deep sequence model, the performance of deep convolutional models that can the visual domain can be improved.

D. *Convolutional neural networks for group activity recognition*

Deng *et.al* proposed a new approach for based on the graphical models for group activity recognition from the surveillance cameras. The group activity recognition is more generalized problem to solve which includes interactions between the human agents with each other as well as the environment and objects in the scene.

The proposed method uses deep networks to recognize the actions of individual people in a scene, and then a neural-networks based hierarchical graphical model refined the predicted labels for each class by considering the dependencies between the classes. The graphical model will be operating as a message-passing similar to other probabilistic graphical models .Figure below shows the general work flow of this approach including the convolutional layers and the message passing step.

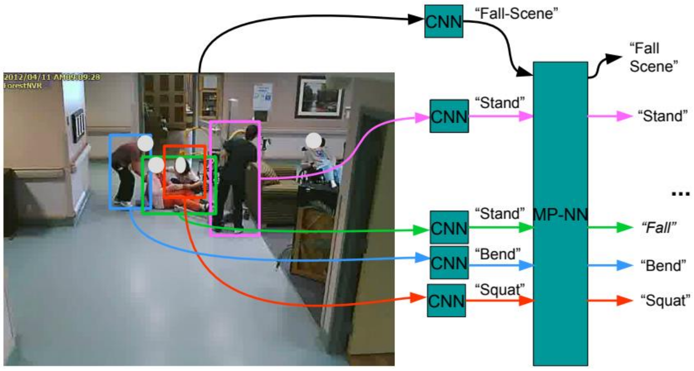

Figure 8. The general workflow of combining graphical models and convolutional neural networks [10]

The proposed architecture by Deng *et.al* consists of two main stages as mentioned earlier. Stage 1 included three different convolutional neural networks namely as Scene ConvNet, Action ConvNet and Pose ConvNet. The Scene ConvNet will predict the label for the whole scene while the action ConvNet will predict the action label for each individual in the scene. The Pose ConvNet will provide the label for the pose of the body for each person in the scene.

The predicted scores from each one these separate networks will be fused with a graphical model. The graphical model learns the message passing parameters and performs inference and learning in a unified framework using the back-propagation method. In fact this method will fuse all the extracted scores from different ConvNets via a graphical model and provides a more general and accurate prediction through a message passing mechanism.

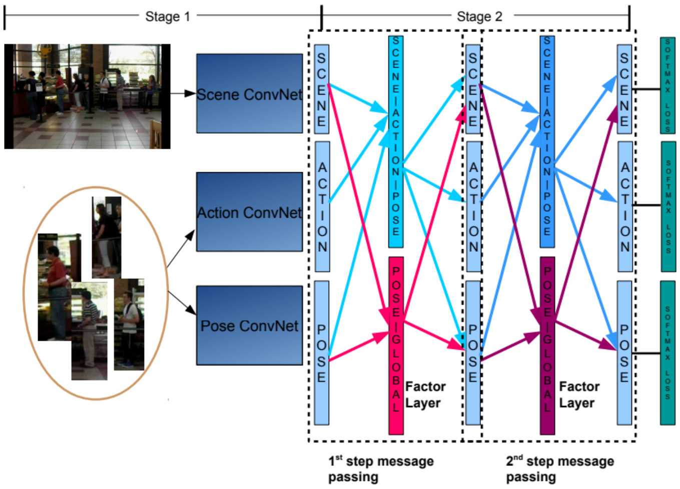

Figure 9. A schematic of the message passing graphical model along with the three ConvNets [10]

IV. Summary and Conclusions

Within this survey paper, an overview of the several approaches for human activity recognition was provided. Before the emerge of deep learning most of the research on the human activity recognition was based on the handcrafted features and pre designed activity recognition pipelines. Within the first section of the paper, a glimpse of these classical techniques in activity recognition was discussed. The next section of the paper was mainly focused on the deep learning approaches for human activity recognition. Deep learning approaches are more recent methodologies due to the large amount of available digital data and increased processing power provided by GPUs. Four different approaches from the literature on the applications of deep learning activity recognition was presented. The first approach is the was using one stream of three-dimensional convolutional neural network for labeling the actions being performed in the video. The one stream approach uses the optical flow, gradient of the frame and the stack of frames as the feature maps to predict the video label.

Two-stream convolutional neural network are also a major family of activity recognition approaches. The idea of two-stream approach is to combine two different ConvNets each responsible for processing a unique portion of information embedded in video. One of the convolutional neural networks will process the temporal domain variations such as the patterns occurring in the optical flow, and the other network will be processing the spatial information at each frame. Then the scores provided by each one of the networks will be fused to conclude the final label for the action.

Other than activity recognition, we also introduce video description methodologies. It has been noted that video description can be a much more challenging task compared to activity recognition. In fact video description generation can be considered as an overlapping area between the computer vision and natural language processing. The techniques being used to solve this problem use convolutional neural network to learn the spatial and temporal patterns and a LSTM network to capture the pattern in the sequence of words to most accurately describe the video.

Predicting the activity that is being performed by a group of people in an environment is another challenging subset of problem in the context of activity recognition. In order to address this problem convolutional neural networks can be also combined with graphical models to provide a better inference regarding the activity recognition. The last approach that has been introduced in the papers includes the data fusion from three different convolutional neural networks via a message-passing stage to predict the group action label.

This survey papers show that how the community of computer vision has been shifting their solutions from classical approaches and handcrafted features towards the deep learning approach. Moreover, in recent year combing different networks has been a practical technique to solve some of the most challenging problems in activity recognition such as video description or group activity recognition.

V. References

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing,* vol. 28, pp. 976-990, 2010.

[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE transactions on pattern analysis and machine intelligence,* vol. 29, pp. 2247-2253, 2007.

[3] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding,* vol. 115, pp. 224-241, 2011.

[4] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 23, pp. 257-267, 2001.

[5] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 35, pp. 221-231, 2013.

[6] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems,* 2014, pp. 568-576.

[7] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," 2016.

[8] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," *arXiv preprint arXiv:1605.03688,* 2016.

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko*, et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625-2634.

[10] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari*, et al.*, "Deep structured models for group activity recognition," *arXiv preprint arXiv:1506.04191*, 2015.