Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Human action recognition using fusion of features for unconstrained video sequences

Chirag I Patel[a,*], Sanjay Garg[a], Tanish Zaveri[b], Asim Banerjee[c], Ripal Patel[d]

[a] *Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, India*
[b] *Electronics & Communication Engineering, Institute of Technology, Nirma University, Ahmedabad, India*
[c] *DA-IICT, Gandhinagar, India*
[d] *BVM Engineering College, Vallabh Vidyanagar, India*

ARTICLE INFO

Article history:
Received 12 September 2015
Revised 10 June 2016
Accepted 13 June 2016
Available online xxx

Keywords:
Human action recognition
Histogram of oriented gradient (HOG)
Local binary pattern (LBP)
Artificial neural network
Support vector machine
Multiple kernel learning
Decision combination neural network (DCNN)
Choquet's fuzzy integral (CFI)
Decison template

ABSTRACT

Effective modeling of the human action using different features is a critical task for human action recognition; hence, the fusion of features concept has been used in our proposed work. By fusing several modalities, features, or classifier decision scores, we present six different fusion models inspired by the early fusion schemes, late fusion schemes, and intermediate fusion schemes. In the first two models, we have utilized early fusion technique. The third and fourth models exploit intermediate fusion techniques. In the fourth model, we confront a kernel-based fusion scheme, which takes advantage of kernel basis of classifiers i.e. Support Vector Machine (SVM). In the fifth and sixth models, we have demonstrated late fusion techniques. The performance of all models is evaluated with ASLAN and UCF11 benchmark dataset of action videos. We obtained significant improvements with the proposed fusion schemes relative to the usual fusion schemes relative state-of-the-art methods.

1. Introduction

In the field of computer vision, a substantial amount of work has been dealt with spatial pattern recognition that involves extracting and identifying objects of interest from video sequences. However, by adding temporal expression, the power of the camera can be drastically increased and used to solve a huge variety of very composite and complex problems. Action recognition refers to an algorithm that the computer system uses to automatically recognize what human action is being or was performed, in a given video sequence. Over the last decade, action recognition has become a crucial research domain for many applications in computer vision field. Actually, it is the problem of classifying an action and assigning it a label of action class. Detecting action performed by humans using camera has a large impact in the industry domain; when a human performs action, his/her body goes through signature movement of body parts. To detect this movement and hence the action performed, computer science researcher needs to design a video system. Most of the recent research works done on action recognition are focused on videos, which are having different constraints, and hence performing action recognition on this video is a challenging task. Recently, these types of video benchmark dataset are released for action video performed by one or more than a human. Still, there is far more distance between the current progress work and real-time action

[*] Corresponding author.

http://dx.doi.org/10.1016/j.compeleceng.2016.06.004
0045-7906/© 2016 Elsevier Ltd. All rights reserved.

recognition. The problem is no consideration of different uncontrolled environments and setting in action video, which seems to happen in real-time action.

The paper is organized as follows. Section 2 describes the brief survey of Human action recognition techniques. Section 3 describes the framework of proposed system. Section 4 includes State-of-the-art comparison with ASLAN and UCF11 benchmark datasets. Finally, the paper is concluded in Section 5.

2. Related work

Local and spatial-level features are gaining much popularity in recent years. Most of the recent works and past works are briefly considered here. Action recognition approaches have been designed to employ high-end applications. Le et al. [1] proposed a new approach for action recognition based on unsupervised feature learning. In this approach, independent subspace network is extended for video to recognize action and features, which are extracted becoming invariant to any change in spatial or temporal axes. This method is simple but the computational cost to fit is high. Another action recognition algorithm is based on one-shot-similarity concept [2]. Three video descriptors, Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOF), and a composition of these two referred to as HNF, are used. Dimensionality is reduced by Principal Component Analysis (PCA). One Shot Similarity (OSS) in the transformed space is defined as OSSML, which further reduces the dimensions of the features. Finally, based on the combination of similarity score, linear SVM classifies actions. Again, this method also has high computational cost as it uses two-dimensionality techniques. Action recognition can be formulated as action similarity problem. Gross et al. [3] proposed a dataset and benchmark for action similarity. In this approach, Space Time Interest Points (STIP) and three other types of local descriptors, HOG, HOF, and composition of these two referred as HNF, are used as features of the system. Linear SVM classifier is used to classify actions. The work presented in [4] mainly focuses on capturing local changes in motion direction for action recognition. The basic strategy used here is motion encoding video using Motion Interchange Pattern (MIP), considering actions, is encoded into a single vector. Linear SVM is used for labeling actions. This method can work on segmented video as well as on unsegmented video. But, this method is not invariant to light changes.

Hanani et al [5] extended the work of [4] for making MIP invariant to illumination; MIP encoding is done for gradient-based description. Two types of encoding techniques are represented in this approach: The first variant encodes each frame in terms of histogram of oriented gradients and the second variant applies MIP after the difference of Gaussian representation. Action recognition problem is formulated as action similarity labeling problem [6]. Two types of feature extraction techniques are used in this approach, which are Fisher vector and vector of locally aggregated descriptor. Both features are having large size; so, Large Margin Dimensionality Reduction (LMDR) method is used. Computational cost is higher for this approach. In LMDR, too many iterations lead to over-fitting, which decreases the performance. In [7], action recognition problem is formulated as a ranking problem. Dense-Scale Invariant Feature Transform (SIFT) and HOG/HOF features are extracted in feature extraction stage. Then, action is recognized using binary ranking model, but creating ranking model seems to be costly. Hassner et al [8] provide a review of action recognition approaches. The author provides a survey of all benchmark datasets related to action recognition. Summary of the results obtained in the last couple of years are also mentioned in this review. Finally, many challenges have been proposed to enhance and modernize the action recognition systems.

Action similarity concept [9] is used to implement action recognition. If a pair of videos is given, the task is to decide if they are same or not. The author proposed an approach, which jointly learns the features and metrics directly from the voxel for measuring action similarity. A set of spatial-temporal cuboids randomly sampled from the video. Auto-encoder is used to encode the clouds. Neural networks classify whether the actions are same or not. Recent action recognition approach is proposed by Siddiqi et al. [10]. Human object is segmented from background by reference frame difference method (where reference frame is an image having no objects) and to override the effect of illumination, histogram of equalization is performed. Feature extraction is executed using wavelet transform decomposition. Feature selection is done using step-wise Linear Discriminant Analysis (LDA). Hidden Markov Model (HMM) is used to classify the features and to recognize the human actions. Another work is presented about action recognition in [11]. Features used in this approach are MIP [4], DogMIP [5], histMIP [5] and dense MBH. To classify actions, word2vec, a simple log-linear classification network, is used.

3. Proposed system description

The framework for the proposed system is shown in Fig. 1. Moving object detection and segmentation, feature extraction, fusion of features, and classifier are the main stages of our system.

3.1. Moving object detection and segmentation

Background is a redundant part in recognizing action because action is related only with the body movements of an object. Moving object is segmented from background using visual attention technique and we have used the approach from [12]. This technique is independent from background model as there is no need of initialization or updation of background model and it is giving good results. The detection of moving objects in [12] follows the procedure detailed below.