Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 430 – 437

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Human Action Recognition based on Spectral Domain Features

Hafiz Imtiaz[a], Upal Mahbub[a], Gerald Schaefer[b], Shao Ying Zhu[c], Md. Atiqur Rahman Ahad[d]

[a]*Bangladesh University of Engineering and Technology, Dhaka, Bangladesh*
[b]*Loughborough University, Loughborough, U.K.*
[c]*Department of Computing and Mathematics, University of Derby, U.K*
[d]*University of Dhaka, Dhaka, Bangladesh*

Abstract

In this paper, we propose a novel approach towards human action recognition using spectral domain feature extraction. Action representations can be considered as image templates, which can be useful for understanding various actions or gestures as well as for recognition and analysis. An action recognition scheme is developed based on extracting spectral features from the frames of a video sequence using the two-dimensional discrete Fourier transform (2D-DFT). The proposed spectral feature selection algorithm offers the advantage of very low feature dimensionality and thus lower computational cost. We show that using frequency domain features enhances the distinguishability of different actions, resulting in high within-class compactness and between-class separability of the extracted features, while certain undesirable phenomena, such as camera movement and change in camera distance, are less severe in the frequency domain. Principal component analysis is performed to further reduce the dimensionality of the feature space. Experimental results on a benchmark action recognition database confirm that our proposed method offers not only computational savings but also a high degree of accuracy.

Peer-review under responsibility of KES International

Keywords: Action recognition; motion representation; 2-D discrete Fourier transform (2D-DFT).

1. Introduction

A human action can be defined as a human body motion that can be described in a non-ambiguous way by one or several verbs. Classification of human actions is a very challenging problem. In order to improve machine capabilities in real-time, e.g., surveillance, human interaction with machines and for helping disabled people, medicine, sports analysis, film, games, augmented reality, etc., it is desirable to represent motion [1]. However, due to various limitations and constraints, no single approach is sufficient for various applications in action understanding and recognition [2].

Present action recognition methods can be classified into view/appearance-based, model-based, space-time volume-based, and direct motion-based methods [3]. Most approaches define an action as a set of local features given by spatio-temporal events or a set of specific human-body poses. Spatio-temporal interest points have been widely successful [4]. A video sequence is represented by a bag of spatio-temporal features called video-words by quantising the extracted 3D interest points (cuboids) from videos, adding a quantised vocabulary of spin-images [4]. Applying scalespace theory, [5] used spatio-temporal interest points that are scale-invariant (both spatially and temporally) and densely cover

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.161

the video content. On the other hand, features based on shape representations have also been extensively investigated, the most notable being shape contexts[6], motion history images (MHIs)[3,7] and space-time shapes[8]. Features based on video volume tensors have also been utilised[9]. Optical flow-based action detection methods are also well-known[10,11,12,13]. For example,[14] recognises human actions at a distance in low-resolution by introducing a motion descriptor based on optical flow measurements. However, this approach cannot deal with large motion such as rapid move across frames. Usually, optical flow is used with other features, because it is noisy and inconsistent between frames[15,16]. Recently, some frequency domain approaches[17] and wavelet-based approaches[18,19] have been shown to offer good recognition accuracy. For example, in[18] wavelets are used for proposing local descriptors utilising the capability in compacting and discriminating data, whereas in[19] wavelet processing techniques are applied to solve the problem of real time processing as well as to filter the original signal in order to achieve better classification.

Unlike methods that use spectral domain features as a means for action recognition, in this paper we propose to extract distinguishable features among different actions to select features from the spectral domain. In our proposed action recognition scheme, a feature extraction algorithm using the two-dimensional discrete Fourier transform (2D-DFT) is developed, which operates within the frames of video sequences to extract features. We show that the discriminating capabilities of the proposed features extracted from the video sequence frames are enhanced because of the spectral-domain feature extraction. Apart from considering only the significant spectral features, further reduction of the feature dimensionality is obtained by employing principal component analysis. Finally, recognition is carried out using a distance based classifier.

2. Proposed method

For any type of recognition, feature extraction is a crucial task which directly dictates the recognition accuracy. For non-stationary and complex backgrounds, it is often difficult to infer the foreground features and the complex dynamics that are related to an action. Moreover, motion blur, occlusions and low resolution present additional challenges that cause the extracted features to be largely noisy. Thus, obtaining an appropriate feature space considering these phenomena for human action recognition is crucial.

2.1. Spectral Feature Selection

In case of frequency domain feature extraction, pixel-by-pixel comparison between action images in the spatial domain is not necessary. Phenomena such as rotation, scale and illumination are more severe in the spatial domain than in the frequency domain. Hence, we intend to develop an efficient feature extraction scheme using 2D-DFT, which offers an ease of implementation in practical applications. For a function $f(x, y)$ with two-dimensional variation, the 2D Fourier transform is given by[20]

$$\mathcal{F}(\omega_x, \omega_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)e^{-j2\pi(\omega_x x+\omega_y y)}dxdy, \tag{1}$$

where ω_x and ω_y represent frequencies in the two-dimensional space.

Our proposed method has two stages: training and classification. The training stage selects features from the motion images resulting from the employment of 2D-DFT, while the classification stage compares an unknown action feature against the set of trained action features. The main concept taken into account by the proposed feature extraction stage is that high amplitude DFT coefficients do concentrate more energy than others. Also, one can notice that it is not true that these high amplitude coefficients are always located in the lower part of the spectrum.

In the training phase, for a given action, f frames are extracted from each one of the q sample video sequences and converted to frequency domain by 2D-DFT. Let us assume that $W = N \times M$ is the number of DFT coefficients from each frame or action image of dimension $N \times M$, and $x_{i,j}$ is the i-th coefficient value of the j-th action image, where $i = 1, 2, 3, ..., W$ and $j = 1, 2, 3, ..., f$. The DFT coefficients obtained from a particular action image are sorted in descending order depending on their amplitudes and the top θ coefficients are selected as distinguishable features for that action image. This step is repeated for all f frames for the particular action. Juxtaposing all the features from all these action images then forms the feature vector for the sample video sequence of the particular action. Therefore, the dimensionality of the feature vector is $1 \times f\theta$. As there are q training sample video sequences, the final feature matrix of a particular action is of dimensionality $q \times f\theta$.