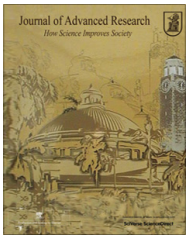
Journal of Advanced Research (2015) 6, 163–169

Cairo University

Journal of Advanced Research



ORIGINAL ARTICLE

An enhanced method for human action recognition

Mona M. Moussa [a,*], Elsayed Hamayed [b], Magda B. Fayek [b], Heba A. El Nemr [a]

[a] *Computers and Systems Department, Electronics Research Institute, Egypt*
[b] *Computer Engineering Department, Faculty of Engineering, Cairo University, Egypt*

ARTICLE INFO

Article history:
Received 28 July 2013
Received in revised form 26 November 2013
Accepted 27 November 2013
Available online 5 December 2013

Keywords:
SIFT
Action recognition
Bag of words
SVM

ABSTRACT

This paper presents a fast and simple method for human action recognition. The proposed technique relies on detecting interest points using SIFT (scale invariant feature transform) from each frame of the video. A fine-tuning step is used here to limit the number of interesting points according to the amount of details. Then the popular approach Bag of Video Words is applied with a new normalization technique. This normalization technique remarkably improves the results. Finally a multi class linear Support Vector Machine (SVM) is utilized for classification. Experiments were conducted on the KTH and Weizmann datasets. The results demonstrate that our approach outperforms most existing methods, achieving accuracy of 97.89% for KTH and 96.66% for Weizmann.

Introduction

Human action recognition is an active area of research due to the wide applications depending on it as detecting certain activities in surveillance video, automatic video indexing and retrieval, and content based video retrieval.

Action representation can be categorized as: flow based approaches [1], spatio-temporal shape template based approaches [2,3], tracking based approaches [4] and interest points based approaches [5]. In flow based approaches optical flow computation is used to describe motion, it is sensitive to noise and cannot reveal the true motions. Spatio-temporal shape template based approaches treat the action recognition problem as a 3D object recognition problem and extracts features from the 3D volume. The extracted features are very huge so the computational cost is unacceptable for real-time applications. Tracking based approaches suffer from the same problems. Interest points based approaches have the advantage of short feature vectors; hence low computational cost. They are widely used and are adopted in this work.

One of the widely used techniques in the action recognition task is Bag of Video Words (BoVW) [6]; which is inspired from bag of words model in natural language processing, where videos are treated as documents and visual features as words [7,8]. This approach proved its robustness to location changes and to noise. Usually the system consists of four main steps: interest-points detection, features description, vector quantization and normalization of the features to construct histogram representation. Finally the histograms are used for classification.

In this work SIFT [9] is used for detecting interest points where the extracted features are invariant to scale, location and orientation changes. 2D SIFT has another advantage which is the limited size of the features vectors; which consumes less computation time than other techniques such as

* Corresponding author. Tel.: +20 233310515.
E-mail address: mona.moussa@gmail.com (M.M. Moussa).
Peer review under responsibility of Cairo University.

Production and hosting by Elsevier

2090-1232 © 2013 Production and hosting by Elsevier B.V. on behalf of Cairo University.
http://dx.doi.org/10.1016/j.jare.2013.11.007

3D descriptors [2,3]. In addition, the accuracy is better than all (to our knowledge) previous work in this field.

The rest of the paper is organized as follows: the next section reviews previous related work, then the proposed system is presented followed by the experiments and results, and finally the conclusion.

Related work

Global descriptors that jointly encode shape and motion were suggested by Lin et al. [10], while Liu and Shah [11] suggested a method to automatically find the optimal number of visual word clusters through maximization of mutual information (MMI) between words and actions. MMI clustering is used after *k*-means to discover a compact representation from the initial codebook of words. They showed some performance improvement.

Bregonzio et al. [12] exploited only the global distribution information of interest points. In particular, holistic features from clouds of interest points accumulated over multiple temporal scales are extracted. A feature fusion method is formulated based on Multiple Kernel Learning.

Chen and Hauptmann [5] proposed MoSIFT which detects interest points then encodes their local appearance and models the local motion. First the well-known SIFT algorithm is applied to find visually distinctive components in the spatial domain and detect spatio-temporal interest points with (temporal) motion constraints. The motion constraint consists of a 'sufficient' amount of optical flow around the distinctive points.

Niebles et al. [13] used probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA) to automatically learn the probability distributions of the spatial–temporal words and the intermediate topics corresponding to human action categories. The system can recognize and localize multiple actions in long and complex video sequences containing multiple motions.

Sadanand and Corso [14] presents a high-level representation of video where individual detectors in this action bank capture example actions, such as "running-left" and "biking-away," and are run at multiple scales over the input video; it represents a video as the collected output of many action detectors that each produces a correlation volume. Being a template-based method, there is actually no training of the individual bank detectors, the detector templates in the bank are selected manually. This method requires using a number of action templates as detectors, which is compositionally expensive in practice.

Tran et al. [15] combined both local and global representations of the human body parts, encoding the relevant motion information as well as being robust to local appearance changes. It represented motion of body parts in a sparse quantized polar space as the activity descriptor.

Fathi and Mori [1] constructed a mid-level motion features built from low-level optical flow information (which is sensitive to noise). These features are focused on local regions of the image sequence, computed on a figure-centric representation, and are created using a variant of AdaBoost. Mid-level shape features were constructed from low-level gradient features using also the AdaBoost algorithm.

Kovashka and Grauman [16] first extract local motion and appearance features from training videos, quantizes them to a visual vocabulary, and then forms candidate neighborhoods consisting of the words associated with nearby points and their orientation with respect to the central interest point. Descriptors for these variable-sized neighborhoods are then recursively mapped to higher-level vocabularies, producing a hierarchy of space–time configurations at successively broader scales.

Methodology

The proposed system is composed of four stages (as shown in Fig. 1): detection of interesting points, feature description for the detected points, building the codebook and finally the classification.

Enhanced interesting points detection

First step in the system is interest points detection where SIFT is utilized to do this process, using algorithm [17]. Fine tuning the threshold parameter is performed to adjust the number of interest points automatically according to the amount of details in each frame. The fine tuning is done by initially apply threshold value = 6 then according to the number of extracted interesting points (np) the threshold (th) is set to a new value as follows:

```
if np > 25 then th = 14
else if np > 20 then th = 10
else if np > 10 then th = 8
else th = 6
```

The threshold value determines the amount of details the detector returns, so when the threshold value is high only the important interest points are detected, while the weak interest points are neglected. Thus the useful information is not lost.

Fig. 2 shows the enhancement achieved by adjusting the threshold. It is obvious that without using a threshold the number of extracted points is very high and they are insignificant where most of them lied in the background. Utilizing a threshold, only the significant points are detected without the need for an additional segmentation step which represents significant processing overhead.

Features description

The SIFT feature vector consists of 128 elements, the coordinates of each point (the x and y location in the frame) are

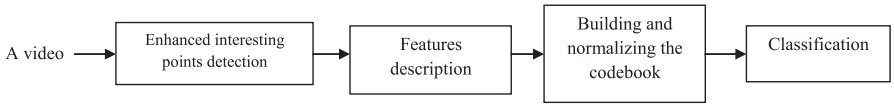


Fig. 1 A block diagram of the proposed system.