J. Vis. Commun. Image R. 33 (2015) 193–202

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

A self-adaptive weighted affinity propagation clustering for key frames extraction on human action recognition ☆,☆☆

Wang Yongxiong *, Sun Shuxin, Ding Xueming

Key Laboratory of Modern Optical System, and Engineering Research Centre of Optical Instrument and System, Ministry of Education, University of Shanghai for Science and Technology, 200093 Shanghai, China

ARTICLE INFO

Article history:
Received 25 January 2015
Accepted 25 September 2015
Available online 1 October 2015

Keywords:
Key frames extraction
3D video sequences
Energy feature
Kinetic energy
Potential energy
Self-adaptive weighted AP clustering
SVM
Human activity recognition

ABSTRACT

In this paper, we propose a novel approach for key frames extraction on human action recognition from 3D video sequences. To represent human actions, an Energy Feature (EF), combining kinetic energy and potential energy, is extracted from 3D video sequences. A Self-adaptive Weighted Affinity Propagation (SWAP) algorithm is then proposed to extract the key frames. Finally, we employ SVM to recognize human actions on the EFs of selected key frames. The experiments show the information including whole action course can be effectively extracted by our method, and we obtain good recognition performance without losing classification accuracy. Moreover, the recognition speed is greatly improved.

1. Introduction

Human action recognition based on computer vision has been widely applied in many domains such as robotics, video surveillance, human computer interaction and content-based video indexing [1–3]. However, with the wide use of camera, there are too many videos or images which need be processed or retrieved by human being or computer. Although many approaches for human action recognition [4,5] have been proposed, most of them perform action recognition by operating on entire video sequences, which requires lots of parameter estimation and large computation time. Schindler and Van Gool [6] found that short action snippets with a few frames (e.g., 1–10 frames) are almost as informative as the entire video. Besides, a few of methods perform human action recognition on key frames [7,8]. All of these show the feasibility and necessity of key frames extraction in action recognition. Thus, in this paper, we propose a novel method to extract key frames for human action recognition.

When people recognize human actions, they always pay more attention to the moving body and informative pose. So we improve this biological idea and introduce the energy features of skeleton joints, including kinetic energy and potential energy, to represent the informativeness of human actions quantitatively and intuitively.

Clustering can identify a subset of representative examples of the original data, which can be applied to the key frames extraction. Nevertheless, traditional clustering algorithms are sensitive to the choice of initial clustering centers. Different from many classic clustering algorithms, Affinity Propagation (AP) makes all the data points as the potential exemplars, which could avoid the initial clustering center influencing of clustering results. Thus, we develop AP algorithm for key-frame extracting.

In this paper, a new method based on key-frame extraction is developed for human action recognition which identifies the frames carrying the most discriminative information. We extract an Energy Feature (EF) for each frame of an action sequence. Since not all the frames are necessarily useful for action recognition, we propose a Self-adaptive Weighted AP has to be (SWAP) clustering algorithm which the weights can be adjusted with the energy feature automatically and employ it to select key frames for each action sequence. Finally, a Support Vector Machine (SVM) [9] is utilized as a classifier for recognizing actions. Intuitively, the run time of classification is largely reduced. The method is significant for the scenarios where highly require real-time performance, for instance, the retrieval of long videos, the pre-processing for manual retrieval, etc. Our method can drastically cut down the search space.

☆ This paper has been recommended for acceptance by M.T. Sun.

☆☆ This work is supported in part by the National Natural Science Foundation of China under Grant 61374039 and 61403254 and Hujiang Foundation of China (C14002, B1402/D1402 and D15009).

* Corresponding author.
E-mail address: wyxiong@usst.edu.cn (Yongxiong Wang).

http://dx.doi.org/10.1016/j.jvcir.2015.09.013
1047-3203/

The main contributions of this paper are summarized as follows: Firstly, an EF is proposed to represent human actions. The EFs consist of kinetic energy features and potential energy features, which can quantitatively represent the information of human actions. Secondly, we propose a SWAP clustering algorithm, and utilize it to select key frames from action video sequence. The proposed SWAP clustering algorithm is compared with the AP clustering and shows significant advantages on key-frame extraction.

The rest of paper is organized as follows: in Section 2, related researches in field of key frames extraction are reviewed. Section 3 provides the details of the proposed framework for key frames extraction, including feature extraction and SWAP clustering algorithm. In Section 4, the experimental results are presented. Section 5 concludes the paper.

2. Related work

Schindler and Van Gool [6] found that not all the frames in a sequence are meaningful to action recognition. In the area of human action recognition, not too many approaches to extract key frames were proposed. Carlsson and Sullivan [10] demonstrated that specific actions can be recognized in single frames. They presented an algorithm for locating action specific frames in video sequences. Their experiments show that using key-frames within the sequence of all forehand strokes are found with no false positives. Hu and Zheng [5] investigated key-frames which are automatically extracted from the image sequence according to the Zernike moment of the silhouette. Thanh et al. [11] extracted the key frames for unit actions based on self-similarity, and labeled them as different patterns. Then they used a statistical metric to evaluate the discriminative capability of each pattern, and defined the bag of reliable patterns as local features for action recognition. Cao et al. [7] have developed the novel approach of key-pose selection which utilizes a PageRank-based centrality measure to select key poses for action recognition. In [12], a benchmarking tool proposed by Lux et al. [13] was used to extract key frames. In such tool key frames are selected by a clustering algorithm based on low-level features. After extracting key-frames, Neda skimmed the video clip by concatenating excerpts around the selected key-frames. This shorter sequence was used as input for SVM classifier. Huang et al. [14] extracted the key frames by using a graph-based method. They regarded each key frame, denoted by both its index and the neighborhood size it represents, as one node in a graph. Edges are only inserted between two nodes that could be neighboring key frames. The weight of each edge was evaluated by the difference between the key frames of its two nodes. They used the shortest path algorithm to find the optimal solution. Liu et al. [8] extract a Pyramidal Motion Feature (PMF) for each frame of an action and use the AdaBoost learning algorithm to select key frames sequence. Finally, they use a SVM classifier for recognizing actions. In [15], energy information of human action including kinetic energy and potential energy were considered as action similarity criterion, the key frames are located at sphere of maximum energy information.

Human actions, from a biological standpoint, could be modeled by the motion of a set of skeleton joints. Most representation methods of human actions based on human skeleton describe positions and velocities of the individual joints, which do not directly show the informativeness of the skeletal configuration of human action [16]. Besides, almost all existing representation methods do not conform to physical principle of human actions and are insensitive to special changes [17]. The kinetic energy and potential energy are developed from velocity and position information respectively [15]. They are more sensitive to tiny and special change of human action.

AP clustering algorithm has been used in searching face image, finding gene exon, searching optimal route, etc. Many improved methods based on AP clustering are proposed since Frey and Dueck proposed AP clustering algorithm. Partition Adaptive Affinity Propagation (PAAP) [18] proposed by Sun and Wang is an extension method of AP. The input similarity matrix is divided into several partitions in PAAP. And values of parameter“preference” can be automatically adjusted, and oscillations or escape are automatically eliminate by adjusting damping factor. Zhang et al. proposed K-AP to generate k clusters [19]. To get a given number of clusters, AP has to be launched many times. While K-AP algorithm introduces a constraint in the process of message passing to exploit the immediate results of k clusters. In [20], Jia et al. designed a density adaptive similarity measure to describe the relations between data points more reasonably. They utilized the dimension reduction method based on spectral graph theory to map the original data points to a low-dimensional Eigen space and proposed a density-adaptive AP clustering algorithm based on spectral dimension reduction. Tang et al. proposed an adaptive clustering method for SAGE data analysis, namely Poisson APS [21]. By incorporating the Poisson statistic character of SAGE data, clustering validation measure is used as a cost function of merging and splitting to adjust the clustering result. Besides, there are some other developed AP clustering algorithms [22–25]. Many methods are improved based on AP clustering algorithm, while none of them take the relationship between the dimensions of data into consideration, which is significant to the similarity matrix. Thus, we propose a SWAP for considering the relationship of data.

3. Methodology

In this study, we focus on the key frames extraction for action sequences captured by 3D camera, and then verify the validity of our method on human action recognition. The overall structure of our method is shown in Fig. 1. Firstly, we extract EF for each frame of action sequences. Each EF consists of kinetic energy and potential energy of human skeleton. A proposed SWAP clustering algorithm is then utilized to select the subset of key frames with the most discriminative information. Finally, we employ a SVM to classify the actions.

3.1. Human action representation

With the extensive development and utilization of 3D camera, more and more useful data for human action recognition can be acquired. Microsoft Kinect [26] is a RGB-D camera which can provide color image and precise depth data. At the same time, it can also synchronously show the human skeleton with 20 joints, as is shown in Fig. 2.

Our biologically inspired feature extraction framework is based on the work of [14]. To obtain the kinetic energy of the joints of human skeleton, we first acquire the x, y, and z coordinates of the joints. Then calculate the velocity and kinetic energy of all joint in a joint set as follows,

$$\begin{aligned} e_{K,t} &= \sum_{i=1}^{n} k_i v_{i,t}^2 = \frac{1}{\Delta t^2}\sum_{i=1}^{n} \boldsymbol{K}^T\left(\left|\boldsymbol{P}_{i,t} - \boldsymbol{P}_{i,t-1}\right|\right)^2 \\ &= \frac{k_i}{\Delta t^2} \sum_{\emptyset(x_{i,t}y_{i,t}z_{i,t})\in F_t} \left((x_{i,t} - x_{i,t-1})^2 + (y_{i,t} - y_{i,t-1})^2 \right. \\ &\quad \left. + (z_{i,t} - z_{i,t-1})^2\right) \end{aligned} \tag{3.1}$$

where $\boldsymbol{\Phi}$ is a set of joints, $e_{K,t}$ denotes the total kinetic energy of all the joints in $\boldsymbol{\Phi}$ at frame F_t, and $v_{i,t}$ is the velocity of ith joint at frame F_t. The subscript i represents ith joint ($i = 1 \ldots n$), $\boldsymbol{P}_{i,t}(x_{i,t}, y_{i,t}, z_{i,t})$ is the position of ith joint at frame F_t, $\boldsymbol{K}$ is a parameter column vector.