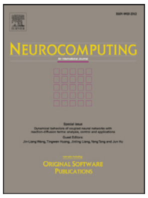
Neurocomputing 273 (2018) 111–119

Contents lists available at ScienceDirect

Neurocomputing

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



Action recognition by Latent Duration Model

Tingwei Wang[a,b], Chuancai Liu[a,*], Liantao Wang[c]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] BS, University of Jinan, Jinan 250013, China
[c] College of Internet of Things Engineering, Hohai University, Changzhou 213022, China

ARTICLE INFO

Article history:
Received 1 January 2017
Revised 8 June 2017
Accepted 25 July 2017
Available online 21 August 2017

Communicated by Dr. Z. Wang

Keywords:
Action recognition
Discriminative segment
Latent duration
Parts-based

ABSTRACT

Temporal structure has attracted lots of research interests due to its ability to model complex pattern for effective action recognition. Most existing methods exploit temporal structure only in a fixed scale or implicit multiple scales. Although recently some methods attempt towards exploiting the temporal structure and relationship by using the durations of action primitives, they cannot effectively conduct the action recognition and discriminative segments discovery simultaneously. In this paper, we propose a novel action recognition method, named Latent Duration Model (LDM), which is a temporal variant of Deformable Part Model (DPM) with explicit durations and temporal ordering constraints. Three types of latent variables are introduced into LDM. Latent duration variables are used to accommodate intra-class temporal scale variation. Latent location variables and latent representation variables are utilized to help search the most discriminative segments in the durations. For temporal structure and relationship, our model takes into account both temporal order and duration changes between consecutive parts, which are robust and flexible to the variety in motion speeds and view angel changes of action videos. Thus, not only discriminative parts with adaptive durations but also robust pairwise relationship is automatically discovered by our model. The experimental results on Olympic Sports, Hollywood2, UCF50 and HMDB51 datasets show the effectiveness of our proposed model.

1. Introduction

Human action recognition from videos is a hot field in the computer vision community [1–3]. It can be applied to intelligent video surveillance, smart home, human-computer interaction and content based information retrieval. Up to date, despite considerable research, action recognition is still an open problem due to the large intra-class variances in illumination changes, view angle changes, camera motions, motion speeds and imaging resolutions.

Bag of words (BoW) based techniques [4], which calculate representation vector by pooling features over the entire video, demonstrate the effectiveness for some simple action data collected from controlled experimental settings. However, BoW has obvious drawbacks brought by ignoring spatial and temporal structure between the words. In fact, a human action can be usually decomposed into a sequence of temporal action primitives. For example, for the action 'drinking', it can be observed that the action primitive 'picks up a glass' is always before the action primitive 'raises the glass'. As such, temporal structure of action primitives provides critical cues for recognizing action. These observations motivate the computer vision community [5,6] including us, to look for temporal patterns in action videos.

In the pipeline of temporal pattern discovery, the popularly employed temporal segmentation methods divide an action video into adjacent segments, each of which corresponds to an informative motion primitive in a short temporal range. Unfortunately, in most cases the action videos to be classified are weakly labeled, which means that none of action primitives, temporal labels, temporal alignment, or spatial-temporal bounding boxes are provided, except for class labels. As a result, it is not straightforward to apply the practiced temporal models such as Conditional Random Field (CRF), Hidden Markov Model (HMM), Dynamic time warping (DTW) for temporal structure modeling.

As study continues, hidden or latent variables are introduced to infer some absent but important information such as low-level spatio-temporal distances between root and discriminative parts (e.g. [7], which is an hierarchical extension of Deformable Part Model (DPM) [8]), high-level semantic context [9] and states of video segments (e.g. Hidden CRF in [6]), etc. For simplicity, it is assumed that a variable is conditionally independent of all other variables given its neighbors, which is called Markov property.

[*] Corresponding author.
E-mail addresses: tingweiwang@outlook.com (T. Wang), chuancailiu@mail.njust.edu.cn (C. Liu), ltwang@hhu.edu.cn (L. Wang).

http://dx.doi.org/10.1016/j.neucom.2017.07.057
0925-2312/© 2017 Elsevier B.V. All rights reserved.

Although these methods are efficient and some of them can realize complex action recognition, there are still some limitations.

(1) It is popular that features are extracted in multiple spatial scales. However, temporal scale, which plays a crucial role for complex action analysis, is not well explored yet. By fixing temporal scale at a time, some methods [6,7,9] hypothesize that all primitives have equal duration and thus fail to exactly model the evolution of primitives in time series.
(2) Existing methods utilize spatio-temporal distances or semantic contexts to describe these pairwise or high-order relationship, which provides important cue to classify different actions. However, these relationship [5,9] are not invariant to temporal scales, and thus susceptible to the changes in motion speeds of and view angles of videos.

To overcome the above limitations, we focus on temporal structure modeling in the context of multiple scales in this paper. Considering the fact that a primitive may have different temporal scales in different action videos according to motion speeds and view angles, our goal can be achieved if temporal scales are appropriately modeled. Inspired by the idea of implicit durations in [10], we explore the temporal scales in terms of durations. In contrast, we model temporal structures among action parts by explicit instead of implicit durations.

The proposed novel action recognition method in this paper is named Latent Duration Model (LDM), which consists of a root and several parts. Part is an equivalent concept with primitive and we will use both of them interchangeably in the following sections if there are no confusions. The durations of action primitives are defined as latent variables since the action data are weakly labeled. To discover the most discriminative segments in the durations, we introduce the other two types of latent variables that correspond to the start position of action primitive and the representative segment in the extent of duration respectively. In addition, we define two types of relationship for robust temporal structure, one is strictly monotoniction for all of the action primitives, and the other is duration ratio for the two consecutive action primitives. For each action class, LDM learns a temporal model of discriminative primitives as well as a classifying boundary in a unified max-margin learning framework. Besides, instead of raw fisher vector, we use PCA to reduce the dimension for segment representation, which still retains comparable recognition accuracies.

The contributions of this paper are highlighted as follows:

(1) Each action primitive is endowed with a variable and latent duration, where the duration allows for intra-class temporal scale variation. Dynamic Programming (DP) is used to exactly infer the best locations and durations of the action primitives. Thus, action recognition can be performed with *multiple temporal scales* for action videos executed in arbitrary motion speed.
(2) We propose a measurement, the ratio of durations, to capture the compatibility between two consecutive primitives. Combining with a strict monotonic property adopted for the temporal order of the action primitives, these ratios are *robust* and *flexible* to measure the temporal structure between primitives in action videos.
(3) For representation of an action part with duration, LDM does not use all of the segments in the duration, while the most *discriminative* one is automatically discovered by DP. Thus, neither aligning of video segment nor sophisticated similarity metric is needed, both of which are very difficult problems.

The rest of this paper is organized as follows: Section 2 reviews the work related to our approach. In Section 3, we present our approach in detail. Experimental results are reported in Section 4. Finally, conclusions are drawn in Section 5.

2. Related work

2.1. Multiple temporal scales

Most of the existing methods deal with scale variation in an implicit and inflexible manner. Methods in [11–14] ignore scale information in model learning phase once discriminative parts, 3D patches or segments candidates with different solutions and temporal lengths are discovered. Spatio-temporal pyramid (STP) representation [15] divides the entire video into sub-volumes with fixed scales and suffers from misalignment caused by different locations of action primitives in various scene layouts. Tian et al. [16] proposed a spatiotemporal DPM, which requires the bounding-box enclosing one cycle of training action, implying that the temporal scales of training samples are provided in advance. Sapienza et al. [17] learned the discriminative space-time action parts from very vast numbers of cuboids with 12 subvolume sizes and a grid spacing of 20 pixels in space and time. Although Raptis et al. [9] utilized a latent scale variable for pairwise term, they did not employ any scale processing to the action part and all pairwise relationship in one sample share the same scale variable.

More recently, several works have explored the idea of modeling temporal scale by hierarchical structure [7,18–20]. However, these hierarchical structures are constructed beforehand and not flexible for intra-class scale variation. For example, Lan et al. [19] performed a sophisticated pre-process for discovering of mid-level action elements and hierarchical trees. Tang et al. [10] proposed a variable-duration hidden Markov model to model durations of states and the transitions between states, where each state has a different variable-duration. In a similar spirit, our LDM also models the durations of action primitives, while it does not suffer from the difficulty of properly inferring probabilities between states and durations as in [10].

2.2. Spatio-temporal relationship

Geometry information, such as location and distance, has been widely utilized for spatio-temporal relationship. For examples, Niebles et al. [5] used a quadratic function of the motion segment displacement only in temporal dimension. [16] extended the displacement function to 3D space. Wang et al. [7] used a function to model the shift of child node from its parent node in a tree-structured model. Raptis et al. [9] exploited the rate of the convergence/divergence to discover the relative motion pattern of two trajectory groups. Moreover, the temporal distribution of the keyposes was modeled using a standard normal distribution in [21].

Other methods learned the compatibility [6] or co-occurrence [22] or transition [20] between a pair of part labels, where indicator functions were used to indicate if the relationship exists in the sample. Sun and Nevatia [23] introduced an evidence localization model where Hidden Markov Model Fisher Vector was used to model temporal consistency. Cheng et al. [24] proposed a high-order segment-level sequence model based on sequence memoizer to explore an infinite length of context in a discrete data sequence.

Temporal ordering constraints have also been utilized for action analysis in [25–28]. Both [27,28] needed more supervised information than class labels: Bojanowski et al. [28] required full time-stamped annotations of elements in the videos, and Nguyen et al. [27] assumed that the change points between actions and the class labels were provided. Raptis and Sigal [25] proposed a jointly learning method for a set of most discriminative keyframes with temporal ordering constraint and local temporal context. Wang and Wu [26] proposed a maximum margin temporal warping method