Image and Vision Computing 46 (2016) 18–28

Contents lists available at ScienceDirect

Image and Vision Computing

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

Local part model for action recognition☆

Feng Shi, Robert Laganière *, Emil Petriu

School of EECS, University of Ottawa, 800 King Edward Ave., Ottawa, ON K1N 6N5, Canada

ARTICLE INFO

Article history:
Received 26 June 2014
Received in revised form 10 August 2015
Accepted 18 November 2015
Available online 16 January 2016

Keywords:
Bag-of-features (BoF)
Action recognition
Random sampling
Local part model
Multi-channel SVM

ABSTRACT

This paper introduces an action recognition system based on a multiscale local part model. This model includes both a coarse primitive level root patch covering local global information and higher resolution overlapping part patches incorporating local structure and temporal relations. Descriptors are then computed over the local part models by applying fast random sampling at very high density. We also improve the recognition performance using a discontinuity-preserving optical flow algorithm. The evaluation shows that the feature dimensions can be reduced by 7/8 through PCA while preserving high accuracy. Our system achieves state-of-the-art results on large challenging realistic datasets, namely, 61.0% on HMDB51, 92.0% on UCF50, 86.6% on UCF101 and 65.3% on Hollywood2.

1. Introduction

The recognition of human actions in videos remains a very active field of research which has a significant impact on a wide range of applications such as intelligent video surveillance, video retrieval, human-computer interaction and smart home systems. Over the last decade, the advances in the area of computer vision and pattern recognition have fuelled a large amount of research with great progress in human action recognition. Much of the early progress [1–3] has been reported on atomic actions with several categories based on staged videos captured under controlled settings, such as KTH [3] and Weizmann [1]. More recently, there are emerging interests for sophisticated algorithms in recognizing actions from realistic video. Such interests involve two prospects: 1) In comparison to image classification evaluating millions of images with over one thousand categories, action recognition is still at its initial stage. It is important to develop reliable, automatic methods which scale to large numbers of action categories captured in realistic settings. 2) With over 100 h of video uploaded to YouTube every minute,[1] and millions of surveillance cameras all over the world, the need for efficient recognition of visual events in videos is crucial for real world applications.

In this paper, we address the problem of fast human action recognition from uncontrolled, realistic video. We propose a solution that achieves both accurate recognition performance and high computational efficiency.

Recent studies [4,5] have shown that low-level local spatio-temporal features and bag-of-features(BoF) can achieve remarkable performance for action recognition on realistic videos. Such approaches have several advantages, such as simplicity, compact video representation, relatively independent representation of events, and better tolerance to illumination, occlusion, deformation and multiple motions *etc.* However, there are still a number of challenges that need to be addressed for action recognition applied in large scale real-world videos.

First, the bag-of-features model only contains statistics of unordered features, and any information related to temporal ordering and spatial structure is lost. In consequence, such approaches have difficulty to discriminate between actions characterized by their structure and event-orderings, such as “stand up” and “sit down”. A more discriminative method should include global structure information and ordering of local events.

Most interest point detectors used for action classification have been extended from the 2D spatial domain. They were originally designed for feature matching, not for selecting the most discriminate patches for classification. Interest point detectors [6] or selected features [7] by unsupervised learning have been shown to be very useful for simple KTH dataset [3] with single, staged human actions and uncorrelated backgrounds. We argue that it is more suitable to include the background information for real-life challenging datasets [8–11] because some of their background features are highly correlated with the foreground actions (*e.g.* diving with water background and skiing with snow background), and thus provide discriminative information for the foreground categories.

It should also be noted that most existing action recognition methods use relatively expensive feature extractor, which could constitute a limiting factor considering the huge amount of data to be processed. In particular, the use of dense trajectories, which is the secret sauce in

☆ This paper has been recommended for acceptance by Ioannis Patras.

* Corresponding author.
E-mail addresses: fshi098@eecs.uottawa.ca (F. Shi), laganier@eecs.uottawa.ca (R. Laganière), petriu@eecs.uottawa.ca (E. Petriu).

[1] http://www.youtube.com/yt/press/statistics.html

http://dx.doi.org/10.1016/j.imavis.2015.11.010
0262-8856/

most state-of-the-art methods, imposes a costly preprocessing step that prevents these methods to be used in real-time scenarios. Moreover, sparse interest point representations may miss important aspects of the scene and therefore do not generate enough relevant information for classification. In contrast, dense sampling methods can provide a very large number of feature patches and thus can potentially produce excellent recognition performance; better results are generally observed as the feature density increases [5,6]. However, the increase in the number of processed points adds to the computational complexity even if simplifying techniques, such as integral video and approximative box-filters, are used.

To overcome these challenges, we proposed a local part model (LPM) to better represent spatio-temporal activities. Our local part model includes both a coarse root ST patch covering local content statistics and finer overlapping part ST patches integrating local structure and temporal relations. To further improve the efficiency of the approach, we use random sampling for feature extraction. An important contribution of this paper resides in the high efficiency of the approach while still producing competitive performances. Our method indeed runs at 30 to 70 fps, depending on the feature used for recognition. This gain in efficiency is achieved by having recourse to two main strategies. First the use of random sampling and integral video for feature extraction. Second by avoiding costly dense trajectory computations and instead relies on global optical flow estimation. We demonstrate in this paper that the use of accurate flow fields is beneficial for action recognition in real-life applications.

The paper is organized as follows: The next section reviews the related works. Section 3 describes the details of our methods. Section 4 introduces different descriptors we used. In Section 5, we present the experimental setup and datasets we tested on. Section 6 summarizes our results and the comparison of our method with other approaches. In terms of recognition accuracy, this is a significant improvement over the state-of-the-art, as well as over our two previous conference publications [12,13]. This paper is built upon these two previous publications. It includes the following additions: 1) an analysis of the impact of dimensionality reduction for efficient action recognition; 2) an improvement of Local Part Model through the use of multiple channels resulting in better performance; 3) an evaluation of the use of more accurate optical flow estimation on performance; and 4) an experimental analysis on different components of the Local Part Model as well as additional experiments on datasets with large numbers of action categories captured in realistic settings. The code to perform random sampling with our Local Part Model is available on-line.[2]

2. Related works

Laptev and Lindeberg [14] were the first to introduce space–time interest point by extending the Harris-Laplace detector to the 3D space. Schüldt et al. [3] built a space–time Harris corner detector with automatic scale selection to detect salient sparse spatio-temporal features. To produce denser space–time feature points, Dollár et al. [2] used a pair of 1D Gabor-filter to convolve with a spatial Gaussian to select local maximal cuboids. Willems et al. [15] proposed the Hessian3D detector and extended the SURF descriptor to detect relatively denser and computationally efficient space–time points. Oshin et al. [16] introduced a Relative Motion Descriptor and used RANSAC to obtain saliency information during interest point detection. Recent works have proposed the use of densely sampled feature points [12,6] and dense trajectories [5,17–19] for action recognition.

Dense sampling has shown to produce good results for image classification [20,21]. For action recognition, Wang et al. demonstrated in [6] that dense sampling at regular space–time grids outperformed state-of-the-art interest point detectors. Similar results have also been observed in [12,17,5]. Compared to interest point detectors, dense sampling generally captures more information by sampling every pixel in each spatial scale. However, such approaches are often computationally intractable for large video datasets.

Uniform random sampling [22], on the other hand, can provide performances comparable to dense sampling. A recent study [23] showed that action recognition performance can be maintained with as little as 30% of the densely detected features. Given the effectiveness of the uniform sampling strategy, one can think of using biased random samplers in order to find more discriminant patches. Yang et al. [24] were able to identify more features on the object of interest by using a prior distribution over patches of different locations and scales. Liu et al. [25] selected the most discriminative subset from densely sampled features using the AdaBoost Algorithm. [26,23] were based on the idea that eye movement of the human viewers is the optimal predictor of visual saliency. They measured the eye movement of human observers watching videos, and used the data to produce an “empirical” saliency map. By using such saliency maps, they pruned 20–50% of the dense features and achieved better results. However, the requirement of prior eye movement data renders such methods impractical for real applications. In addition, because of computational constraints, these methods didn't explore high sampling density schemes to improve their performance.

To deal with the “out-of-ordering” problem of the bag-of-features representation, Hamid et al. [27] proposed an unsupervised method for detecting anomalous activities by using bags of event n-grams. In their method, human activities were represented as overlapping n-Grams of actions. While overlapping n-grams can preserve the temporal order information of events, it causes the dimensionality of the space to grow exponentially as n increases. Thurau and Hlavác [28] introduced n-grams of primitive level motion features for action recognition. Laptev et al. [4] extended image representation of spatial pyramid [29] to the spatio-temporal domain. The authors divided a video into a grid of coarse spatio-temporal cells. The whole video was then represented by the ordered concatenation of the per-cell BoF models. Such ordered concatenation adds global structural information. Gaidon et al. [30] focused on explicitly capturing the spatial and temporal structure of actions with structure model. Tang et al. [31] used a variable-length discriminative HMM model which infers latent sub-actions to explicitly model the presence of sub-events.

As for real-time action recognition algorithms, both Ke et al. [32] and Willems et al. [15] used approximative box-filter operations and integral video structure to speed-up feature extraction. Patron-Perez and Reid [33] employed a sliding temporal window within the video and used first-order dependencies to effectively approximate joint distribution over feature observations given a particular action. Yeffet and Wolf [34] efficiently classified the actions with Local Binary Patterns and an approximated linear SVM classifier. Yu et al. [35] extended the efficient 2D FAST corner detector to the 3D domain V-FAST detector, and applied semantic texton forests for fast visual codeword generation. Whiten et al. [36] exploited very efficient binary bag-of-features matching with the Hamming distance rather than the Euclidean distance through an extension of the popular 2D binary FREAK descriptor.

3. Action recognition using LPM

Nowak et al. [22] have shown that the most important factor impacting a system's performance is the number of sampled patches. While the performance of dense sampling is improved as the sampling step size decreases [5], such approach becomes rapidly computationally intractable due to the very large number of patches produced. To achieve both computational efficiency and high accuracy, our approach increases the sampling density by decreasing the sampling step size, and at the same time controls the number of sampled patches used by the classifier. We experimentally found that, with proper sampling

[2] http://www.site.uottawa.ca/laganier/projects/actionLPM/index.html