Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Weakly supervised activity analysis with spatio-temporal localisation

Feng Gu [a,b,*], Muralikrishna Sridhar [a], Anthony Cohn [a], David Hogg [a], Francisco Flórez-Revuelta [b], Dorothy Monekosso [c], Paolo Remagnino [b]

[a] School of Computing, University of Leeds, LS2 9JT, UK
[b] Faculty of Science, Engineering and Computing, Kingston University, KT1 2EE, UK
[c] School of Computing, Creative Technologies & Engineering, Leeds Beckett University, LS1 3HE, UK

ARTICLE INFO

Article history:
Received 17 July 2015
Received in revised form
15 July 2016
Accepted 11 August 2016
Communicated by Prof. Zidong Wang

Keywords:
Human activity analysis
Spatio-temporal localisation
Weakly labelled video data
Multi-instance multi-label learning

ABSTRACT

In computer vision, an increasing number of weakly annotated videos have become available, due to the fact it is often difficult and time consuming to annotate all the details in the videos collected. Learning methods that analyse human activities in weakly annotated video data have gained great interest in recent years. They are categorised as "weakly supervised learning", and usually form a multi-instance multi-label (MIML) learning problem. In addition to the commonly known difficulties of MIML learning, i.e. ambiguities in instances and labels, a weakly supervised method also has to cope with large data size, high dimensionality, and a large proportion of noisy examples usually found in video data. In this work, we propose a novel learning framework that iteratively optimises over a scalable MIML model and an instance selection process incorporating pairwise spatio-temporal smoothing during training. Such learned knowledge is then generalised to testing via a noise removal process based on the support vector data description algorithm. According to the experiments on three challenging benchmark video datasets, the proposed framework yields a more discriminative MIML model and less noisy training and testing data, and thus improves the system performance. It outperforms the state-of-the-art weakly supervised and even fully supervised approaches in the literature, in terms of annotating and detecting actions of a single person and interactions between a pair of people.

1. Introduction

The annotation and detection of human activities have become increasingly significant research problems in the field computer vision, due to the growing demand of analysing large quantities of available videos. However, the proliferation of videos is often unmatched by the availability of detailed spatio-temporal annotation of activities in these videos, mainly due to the laborious nature of such an effort. Thus, much of the annotation comes in a weakly labelled form, where several class labels are simultaneously provided for a single data unit, i.e. a video, without any information about the spatio-temporal locations of the activities. As a consequence, two types of ambiguities result from such a weakly labelled annotation, making it hard to directly apply conventional supervised learning techniques. The first ambiguity is in the instance location, wherein spatio-temporal locations of the true instances that may correspond to activities in a video are not known a priori for training. The second ambiguity is in the instance label, multiple class labels may be associated with a video, while the true label of each individual instance in the video is not known a priori. The ambiguities of instance location and instance label constitute a weakly supervised learning problem, known as multi-instance multi-label (MIML) learning.

Numerous MIML learning techniques have emerged and formed a powerful weakly supervised learning framework, which is capable of simultaneously dealing with instance location and instance label ambiguities mentioned above. These techniques have been applied to various image datasets [38,30,33], but barely to videos. Similar to the applications to images, we expect that MIML learning would address the instance location ambiguity by generating multiple instances at different spatio-temporal locations in a video in the form of a bag, which is labelled with respect to the class labels given for the entire video. Then it learns to identify the true instances that correspond to real activities in the video. Additionally, MIML learning would resolve the instance label ambiguity by explicitly modelling interclass correlations and tries to learn the true label of each instance that corresponds to one of the activity classes in the video. While investigating a recently introduced MIML model [33] that we found scalable to video datasets, we have observed that during both training and testing, a large number of noisy samples, completely irrelevant to any activities of interest, tend to have an adverse effect on the

* Corresponding author at: School of Computing, University of Leeds, LS2 9JT, UK.
E-mail address: fengy1982@gmail.com (F. Gu).

http://dx.doi.org/10.1016/j.neucom.2016.08.032
0925-2312/

model's learning ability. This problem has been recently studied on various datasets for training multi-instance learners using an approach known as instance selection [6,11].

In this work, we use a scalable MIML model [33] as the base classifier, and incorporate an innovative spatio-temporal smoothing based instance selection process for the purpose of reducing noise in video data. This forms a novel MIML learning framework for annotating and detecting human activities in weakly labelled video data, where the labelling merely provides the presence of activities in each video but not their spatio-temporal locations. Our contributions can be summarised as follows:

(1) An instance selection process is introduced to enforce spatio-temporal smoothing at the bag level, along with the instance classification by the base MIML classifier at the instance level, which is formulated as an energy function similar to the one defined in the minimisation problem of Markov random fields (MRF) [17].
(2) A two-step optimisation is applied to alternate iteratively between the base MIML classifier and the instance selection process, aimed at minimising the MRF like energy function until it converges, which provides the knowledge to distinguish the prototype instances potentially associated with the classes of interest from the noisy ones.
(3) The learned knowledge of instance selection is then generalised to testing via a noise removal process based on the support vector data description (SVDD) algorithm [28], which learns a description of the prototype instances, to identify noisy instances as outliers during testing.

On application to three benchmark video datasets, we have found that the proposed framework significantly improves the performance in terms of the annotation task (to recognise activities and annotate their spatio-temporal locations in a training video) and the detection task (to recognise activities and detect their spatio-temporal locations in a testing video). The results also suggest that it outperforms the original MIML model [33], the state-of-the-art weakly supervised approaches [25,24,18], as well as fully supervised methods [4,29,22,32] in the literature across the three datasets.

The paper is organised as follows: Section 2 provides a review of related work for MIML techniques and weakly supervised action detection; Section 3 details the feature representation of video under the weakly supervised setting; Section 4 formulates the proposed framework and introduces the generation of instances and bags in the setting of weakly supervised action detection; Section 5 describes the experiments, such as data and implementation details; Section 6 demonstrates results and analysis of the experiments; finally Section 7 concludes this work and points out possible future work.

2. Related work

Multi-instance learning and multi-label learning have evolved as two separate paradigms until recently [39], where the authors proposed two solutions to bridge them for the MIML learning problem. The first solution transforms each bag of instances into a single instance and then performs multi-label learning. The second solution generalises a multi-instance single label learning algorithm to handle multiple labels. Subsequently, Zha et al. [38] proposed an undirected graphical model for image classification, which simultaneously captures both the connections between class labels and regions (instances), and the correlations among the labels in a single formulation. Its learning and inference process relies on an expectation maximisation algorithm and approximation methods, e.g. the contrastive divergence algorithm [13] and Gibbs sampling [12], which tend to be slow for problems with a large number of instances. In [30], the authors proposed an active learning framework for image annotation that first divides the multi-label problem into a set of binary classification problems and then devises a multi-label set kernel to weight each instance for the multi-instance learning. This framework exhibits limitations when applying to complex datasets, due to its combined polynomial complexity of labels and instances respectively. The methods above are merely designed for the recognition objects in images without any localisation of recognised objects. Therefore they are not directly applicable to more complex video data, for annotating and detecting human activities, with spatio-temporal localisation of recognised activities.

Hu et al. [14] proposed a multi-instance learning framework, SMILE-SVM, to handle ambiguities in the locations of single person's actions in videos of complex scenes. The framework however relies on manually annotated rough locations of an action in a video for training, and it assumes that a video at most contains one true instance of one of the action classes. Therefore, it cannot be directly applied to weakly labelled video data that provides the presence of multiple activities without any spatio-temporal localisation in each video. A multi-instance learning approach that optimises intraclass and interclass distances for action annotation and detection is introduced in [25]. The training does not require the manual annotation of rough locations of actions, but it still has the same assumption of one true instance of one of the action classes per video. It is optimised by a genetic algorithm, which is known for its slow convergence rate [15]. It is then extended and improved for weakly supervised annotation in [24], by focusing on negative mining in multi-instance learning. Neither of the approaches [24,25] however explicitly models the interclass relationships within each video (or bag), and thus may struggle in cases where multiple action classes are simultaneously presented in a video (i.e. multi-label learning).

A MIML approach [33] is recently introduced for bag level object recognition in images that feature ambiguities in both the instance location and instance label. The approach has been shown highly scalable to the amount of instances, the dimensionality of feature space, and the number of label classes, which is ideal for complex video data. It trains a set of discriminative multi-instance classifiers and models the interclass correlation among labels by finding a low rank weight matrix. This enforces the classifiers to share weights and perform multi-label learning. This approach however is designed for the classification of a bag rather than each individual instance in the bag. It might not be able to distinguish the positive instances associated with the label classes of interest from the rest in a bag, for the purpose of spatial localisation in images or spatio-temporal localisation in videos. The performance could further deteriorate under the influence of noise, e.g. problems with a low signal-to-noise-ratio (SNR), particularly common in video data. As a result, some means of removing noise from the data, especially those instances completely irrelevant to any label classes of interest, would be beneficial. Recent approaches [6,11] on instance selection represent the target concept using multiple prototypes that are formed and updated iteratively, thereby simultaneously eliminating many noisy samples in each bag. While instance selection has shown to be a promising direction, it has so far been applied only to multi-instance learning problems but not MIML ones, and has not been extended to testing. This leads to the motivation of this work, that is, to develop a MIML learning framework that is capable of reducing noise from both the training and testing data through instance selection. Such a framework will be applied and evaluated on weakly labelled video data, for purpose of annotating and detecting human activities with spatio-temporal localisation.

Please cite this article as: F. Gu, et al., Weakly supervised activity analysis with spatio-temporal localisation, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.08.032