Computer Vision and Image Understanding 000 (2017) 1–13

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Image and video mining through online learning

Andrew Gilbert*, Richard Bowden

CVSSP, University of Surrey, GU2 7XH Guildford, United Kingdom

ARTICLE INFO

Article history:
Received 5 September 2016
Revised 7 December 2016
Accepted 1 February 2017
Available online xxx

Keywords:
Action recognition
Data mining
Real-time
Learning
Spatio-temporal
Clustering

ABSTRACT

Within the field of image and video recognition, the traditional approach is a dataset split into fixed training and test partitions. However, the labelling of the training set is time-consuming, especially as datasets grow in size and complexity. Furthermore, this approach is not applicable to the home user, who wants to intuitively group their media without tirelessly labelling the content. Consequently, we propose a solution similar in nature to an active learning paradigm, where a small subset of media is labelled as semantically belonging to the same class, and machine learning is then used to pull this and other related content together in the feature space. Our interactive approach is able to iteratively cluster classes of images and video. We reformulate it in an online learning framework and demonstrate competitive performance to batch learning approaches using only a fraction of the labelled data. Our approach is based around the concept of an image signature which, unlike a standard bag of words model, can express co-occurrence statistics as well as symbol frequency. We efficiently compute metric distances between signatures despite their inherent high dimensionality and provide discriminative feature selection, to allow common and distinctive elements to be identified from a small set of user labelled examples. These elements are then accentuated in the image signature to increase similarity between examples and pull correct classes together. By repeating this process in an online learning framework, the accuracy of similarity increases dramatically despite labelling only a few training examples. To demonstrate that the approach is agnostic to media type and features used, we evaluate on three image datasets (15 scene, Caltech101 and FG-NET), a mixed text and image dataset (ImageTag), a dataset used in active learning (Iris) and on three action recognition datasets (UCF11, KTH and Hollywood2). On the UCF11 video dataset, the accuracy is 86.7% despite using only 90 labelled examples from a dataset of over 1200 videos, instead of the standard 1122 training videos. The approach is both scalable and efficient, with a single iteration over the full UCF11 dataset of around 1200 videos taking approximately 1 min on a standard desktop machine.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Fuelled by the prevalence of cameras on mobile devices and social networking sites such as Facebook, Twitter and YouTube, digital content is ever increasing. This produces a demand for automatic approaches to clustering media into meaningful semantic groups to facilitate browsing and search. This use case is incompatible with traditional supervised training methods, as labelling the data is the limiting factor. Therefore, we propose an approach that allows the user to find natural groups of similar content based on a small handful of seed examples. Combining these seed examples with an automatic data mining approach that extracts rules that can generalise and further cluster the remaining unseen media.

There have been many approaches that are successful in the classification of images and videos (Gilbert et al., 2009; Han et al., 2009; Marszalek et al., 2009; Schuldt et al., 2004; Wang et al., 2009). However, these require significant amounts of supervised training data, which is increasingly infeasible to provide. There are single shot approaches that take a limited training set (Ning et al., 2009; Shechtman and Irani, 2007). However, they can be sensitive to noise in the training data, and are difficult to generalise to larger datasets.

Conversely, we use an online learning approach capable of incrementally clustering similar material from the manual identification of a few correct and incorrect examples. These examples are then used to learn rules that can be applied to clustering a larger corpus of material. The approach is demonstrated on three pure image datasets (15 Scene Lazebnik et al., 2006, Caltech101 Fei-Fei et al., 2004, FG-NET Panis et al., 2015), on a combined text and image dataset (ImageTag Gilbert and Bowden, 2012), a dataset used in active learning (Iris Lichman, 2013) and on three state-of-the-

* Corresponding author
E-mail addresses: a.gilbert@surrey.ac.uk (A. Gilbert), R.Bowden@surrey.ac.uk (R. Bowden).

http://dx.doi.org/10.1016/j.cviu.2017.02.001
1077-3142/© 2017 Elsevier Inc. All rights reserved.

art video action recognition datasets (UCF11 Liu et al., 2009, KTH Schuldt et al., 2004, and Hollywood2 Marszalek et al., 2009).

To provide both scalability and incremental learning, the approach needs to remain efficient as datasets become larger. Therefore, we efficiently compute both distances between high dimensional representations and dynamically augment the representation with new compound elements to form an image signature. We demonstrate the approach is independent of the underlying features. The similarity measure employed in this paper extends the original min-Hash algorithm that was designed to identify the similarity between text in documents (Brooder, 1998) by efficiently computing the distances between high-dimensional sets. Chum et al. (2007) demonstrated the ability of min-Hash to efficiently identify near duplicate images within datasets. Min-Hash is ideally suited to large high dimensional representations, as the computational costs are not proportional to the size of the input representation. This makes it especially suited to complex image or video descriptors which are typically of high dimensionality. Chum et al. (2008) later extended this work to approximate the histogram intersection of images.

Another data mining tool employed in this work is association rule mining (known as APriori Agrawal and Srikant, 1994). This was originally designed to identify co-occurring elements in large text files. It was first employed in the image domain by Quack et al. (2007). They used association rule mining in supervised object recognition to find spatially grouped SIFT descriptors.

In the temporal domain, Gilbert et al. (2009) demonstrated the use of APriori in Action recognition. They argued that many other action recognition approaches (Dollar et al., 2005; Laptev and Lindeberg, 2003; Laptev et al., 2008), use features engineered to fire sparsely, to ensure that the overall problem is tractable. However, they suggested that this can sacrifice recognition accuracy as it cannot be assumed that the optimum features for class discrimination are obtained from this approach. In contrast, an over complete set of Harris corners (Harris and Stephens, 1988) are grouped spatially and temporally, mining is then used to identify feature combinations to classify video sequences. While this demonstrated the power of APriori in activity recognition, the training was still performed with comprehensive supervised training sets.

2. Related works

There is a number of related works that aim to reduce the labeleing of the training data. An online incremental algorithm (such as Law et al., 2004) can reduce the training examples and time required, we propose to include both correct and incorrect instances in a human led iterative process to select fewer but more relevant training examples. As with any approach that clusters or correlates images and video, the choice of the representation and similarity measure is critical, as they can affect both the size of the database and the search time. We introduce the image signature as an efficient representation irrespective of the type of the input sample: image or video or the feature descriptor applied. Then, using APriori, the distinctive and discriminative elements of these selected examples are identified and accentuated across the dataset by dynamically augmenting the representation with new compound elements. This increases the set overlap of correct image signatures while also improving the dissimilarity of incorrectly classified examples thereby increasing the overall accuracy of matching. As the image signature increases in dimensionality, min-Hash provides a scalable approach to computing similarity between data items. This iterative procedure can be seen as a form of online learning with similarities to approaches in both active and metric learning.

Tong proposed active learning for the purpose of image retrieval (Tong, 2001). Active learning is a particular case of semi-supervised machine learning where the learning algorithm interactively queries the user to obtain the desired outputs for new data points. Since the learner can identify examples of great confusion or variation to focus on, the number of examples to label for a concept can often be much lower than the number required in batch. This is a key aspect of our approach, in classical active learning, the algorithm chooses the data points to be labelled based on some automated criteria. Our approach uses the notion of similarity and allows the user to select obvious outliers that should be labelled. Similarity helps the user prioritise annotation, and the feature representation is manipulated to satisfy these constraints. This changes the topology of the distance space and is therefore also related to Metric Learning. Metric learning is the task of learning a distance function over a dataset usually pairwise metric distances between samples.

There have recent developments involving users in hybrid active learning approaches (Weigl and Radauer, 2016; Weigl and Eitzinger, 2016; Lughofer, 2012). Lughofer (2012) employs sample selection in the first phase based purely on unsupervised criteria. Then in the second phase, the task is to update the pre-trained classifiers with the most relevant samples. We propose a similar ideology however allow the user to select the relevant samples via a Multi-Dimensional Scaling (MDS) visualisation and unsupervised clustering of the distance between all data samples together with the novel approach identification of the discriminative features. While Weigl and Eitzinger (2016) is similar to this work through allowing the user to select the most relevant samples based on a visualization map showing the sample/class distributions. However we propose a more generic feature type to ensure multiple data models can be incorporated in this single method. Weigl and Radauer (2016) performs on-line image classification tasks, in this case for event type classification, presenting the user "questionable" events for the user to examine instead of the whole dataset. Although the speed and ease of visualisation and the feature learning within this approach allows the full datasets to be presented to the user at iteration to ensure they don't get stuck in a local minima in the dataset.

2.1. Paper overview

In this manuscript, we build upon our previous work in Gilbert and Bowden (2011a, 2011b) which introduced the online learning framework and was combined with a hand gesture estimation controller (Krejov and Bowden, 2014). This manuscript provides a mature and a detailed description of the approach. We have reformulated the learning framework and provide an extensive formalisation of the method to allow for repeatability. Regarding analysis, additional features have been added and evaluated on seven different datasets, which include a broad range of various modalities (i.e. image, video and combined image/text-tag) using multiple user runs. We also provide analysis regarding cluster purity and evaluation of the computational cost of the approach, showing that the online learning framework can compete favourably with the state of the art supervised learning approaches using only a fraction of the data.

Section 2 introduces the image signature and extends the min-Hash algorithm for video similarity in Section 3. An image signature is a symbolised vector suitable for use by frequency based mining algorithms. The process of symbolisation takes a fixed dimensionality vector, such as a histogram, and converts it into a variable length set of discrete symbols. Each symbol represents a dimension in the original vector, the number of times each symbol appears relates to the magnitude of that dimension. The learning framework is described with clustering and visualisation discussed in Section 4. Section 5 illustrates how frequent itemset mining can be modified to identify discriminative or common elements of the signatures, that are then accentuated (Section 6) to change the

Please cite this article as: A. Gilbert, R. Bowden, Image and video mining through online learning, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.02.001

Download English Version:

https://daneshyari.com/en/article/4968789

Download Persian Version:

https://daneshyari.com/article/4968789

Daneshyari.com