ARTICLE IN PRESS

Neurocomputing xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

Training-free indexing refinement for visual media via multi-semantics

Peng Wang[a,*], Lifeng Sun[a], Shiqiang Yang[a], Alan F. Smeaton[b]

[a] *National Laboratory for Information Science and Technology, Department of Computer Science and Technology Tsinghua University, Beijing 100084, China*

[b] *Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland*

ARTICLE INFO

Keywords:
Semantic indexing
Refinement
Concept detection enhancement
Context fusion
Factorization
Propagation

ABSTRACT

Indexing of visual media based on content analysis has now moved beyond using individual concept detectors and there is now a focus on combining concepts by post-processing the outputs of individual concept detection. Due to the limitations and availability of training corpora which are usually sparsely and imprecisely labeled with concept groundtruth, training-based refinement methods for semantic indexing of visual media suffer in correctly capturing relationships between concepts, including co-occurrence and ontological relationships. In contrast to training-dependent methods which dominate this field, this paper presents a training-free refinement (TFR) algorithm for enhancing semantic indexing of visual media based purely on concept detection results, making the refinement of initial concept detections based on semantic enhancement, practical and flexible. This is achieved using what can be called multi-semantics, factoring in semantics from multiple sources. In the case of this paper, global and temporal neighbourhood information inferred from the original concept detections in terms of weighted non-negative matrix factorization and neighbourhood-based graph propagation are both used in the refinement of semantics. Furthermore, any available ontological concept relationships among concepts can also be integrated into this model as an additional source of external *a priori* knowledge. Extended experiments on two heterogeneous datasets, images from wearable cameras and videos from TRECVid, demonstrate the efficacy of the proposed TFR solution.

1. Introduction

Video in digital format is now in widespread use in everyday scenarios. While mainstream consumer-based access to image and video on platforms such as YouTube and Vine are based on user tags and metadata, prevailing methods to indexing based on *content* detect the presence or absence of semantic concepts which might be general (e.g., *indoor*, *face*) or more abstract (e.g., *violence*, *meeting*). The conventional approach to content-based indexing of visual media, as taken in the annual TRECVid benchmarking [21,20], is to manually annotate a collection of visual media covering both positive and negative examples, for the presence of each concept. This can be done manually, or can use visual captchas [16], and then train a machine learning classifier using these annotations to recognise the presence, or absence, of the semantic concept. This typically requires a classifier for each concept without considering inter-concept relationships or dependencies yet in reality, many concept pairs and triples are often semantically related and dependent and thus will co-occur rather than occur independently. It is widely accepted and it is intuitive that detection accuracy for concepts can be improved if concept correlation can be exploited.

The idea of refining an initial, raw, set of concept detections is intuitive and has been explored for some time and it is still currently a topic attracting a lot of attention, such as in [14]. Context-Based Concept Fusion (CBCF) is an approach to refining the detection results for independent concepts by modeling relationships between them [5]. Concept correlations are either learned from annotation sets [10,24,25,8,6] or inferred from pre-constructed knowledge bases [28,9] such as WordNet. However, annotation sets are almost always inadequate for learning correlations due to their limited sizes and the annotation having being done with independent concepts rather than correlations in mind. In addition, training sets may not be fully labeled or may be noisy. The use of external knowledge networks also limits the flexibility of CBCF because it uses a static lexicon which is costly to create and even costlier to maintain. When concepts do not exist in an ontology, these methods cannot adapt to such situations.

In this paper we propose a training-free refinement (TFR) method to exploit inherent co-occurrence patterns for concepts which exist in testing sets, exempt from the restrictions of training corpus and external knowledge structures and we use this to refine and improve

[*] Corresponding author.
E-mail addresses: pwang@tsinghua.edu.cn (P. Wang), sunlf@tsinghua.edu.cn (L. Sun), yangshq@tsinghua.edu.cn (S. Yang), alan.smeaton@dcu.ie (A.F. Smeaton).

http://dx.doi.org/10.1016/j.neucom.2016.08.107
Received 13 March 2016; Received in revised form 21 June 2016; Accepted 8 August 2016
Available online xxxx
0925-2312/ © 2016 Elsevier B.V. All rights reserved.

the output of independent concept classifiers. TFR can fully exploit various sources of semantic information including global patterns of multi-concept appearance, an ontology encapsulating any concept relations (if available), as well as sampling the distribution of concept occurrences in the temporal neighbourhood of a given image, all with the goal to enhance the original one-per-class concept detectors and all done within a unified framework. Although this reduces the learning/training process, we set out here to see if TFR can still obtain better or comparable performance than the state-of-the-art as such an investigation into refinement of semantic indexing has not been done before.

The contributions of this paper can be highlighted as:

- A training-free refinement method which uses information inferred from test datasets without any requirement for high quality training data based on full concept annotations. This can flexibly adapt to many real world applications where only limited or incomplete annotations are available for correlation inference and goes beyond the state-of-the-art in that it is flexible and dynamically adaptable to new domains or datasets, without the need for a training phase;
- An ontological factorization algorithm to adjust and improve on the initial less accurate results for concept detection, according to the global patterns of concept appearance and absence, across the whole collection of samples. Ontology-based concept relationships can also be combined into this algorithm as another source of external *a priori* knowledge thus illustrating how the TFR method presented here, can easily incorporate new sources of evidence for concept refinement, unlike other available approaches;
- A similarity graph of nearest neighbours based on the refined results using ontological factorization and applying a graph propagation algorithm to further enhance the detection accuracy exploiting such local relationships, which finally achieves satisfactory refinement, something which has not been available previously;
- A set of experiments on two heterogeneous datasets, chosen to validate the effectiveness of the above.

The rest of the paper is organized as follows: in Section 2 we review related work on refinement of semantic indexing. In Section 3 we present an overview of our TFR solution and algorithm followed by a detailed elaboration of TFR in Section 4. A set of experiments including a description of the two datasets we used and a discussion of results, are presented in Section 5. We finish with conclusions and proposals for future work.

2. Related work

The task of automatically determining the presence or absence of a semantic concept in an image or a video shot (or a keyframe) has been the subject of at least a decade of intensive research. The earliest approaches treated the detection of each semantic concept as a process independent of the detection of other concepts and used supervised learning approaches to implement this, but it was quickly realised that such an approach is not scalable to large numbers of concepts, and does not take advantage of inter-concept relationships. Based on this realisation, there have been efforts within the multimedia retrieval community focusing on utilization of inter-concept relationships to enhance detection performances, which can be categorized into two paradigms: multi-label training and detection refinement or adjustment.

In contrast to isolated concept detectors, *multi-label training* tries to classify concepts and to model correlations between them, simultaneously. A typical multi-label training method is presented in [18], in which concept correlations are modeled in the classification model using Gibbs random fields. Similar multi-label training methods can be found in [30]. Since all concepts are learned from one integrated model, one shortcoming is the lack of flexibility, which means that the learning stage needs to be repeated when the concept lexicon is changed. Another disadvantage is the high complexity when modeling pairwise correlations in the learning stage. This also hampers the ability to scale up to large-scale sets of concepts and to complex concept inter-relationships.

There has also been some work on *multi-label detection*, within the framework of TRECVid where for the 2012 and 2013 edition of the TRECVid semantic indexing task, a secondary “concept pair” task was offered. The motivation here is a video (but could equally well be image) retrieval scenario which demands complex queries that go beyond a single concept. Examples of concept pairs which could go together include $Animal + Snow$, $Person + Underwater$ and $Boat/Ship + Bridges$. Rather than combining concept detectors at query time, the TRECVid concept pair task aimed at detecting the simultaneous occurrence of a pair of unrelated concepts in a video.

In 2012 the top run achieved a score of 0.076 *MAP* and in 2013 the top run achieved a score of 0.162 *MAP* [2]. While this seems an improvement, it should be noted that the pairs changed from one year to the next and some may have been easier, or less rare, than the ones in 2012. Of course there was variability in performance across concept pairs but the best performer for the pair $GovernmentLeader + Flags$, for example, scored 0.658 *MAP* which is very respectable.

The approaches taken by various participants in this activity were mostly based around combining multiple individual detectors by well known fusion schemes, including sum, product and geometric mean and while it represents an interesting exploration, the feasibility of indexing visual media, at indexing time, by concept pairs and scaling this to large collections would seem remote.

As an alternative to concept detection at indexing time, *detection refinement or adjustment* methods post-process detection scores obtained from individual detectors, allowing independent and specialized classification techniques to be leveraged for each concept. Detection refinement has attracted interest based on exploiting concept correlations inferred from annotation sets [10,24,25,5] or from pre-constructed knowledge bases [28,9,12]. However, these depend on training data or external knowledge. When concepts do not exist in the lexicon ontology or when extra annotation sets are insufficient for correlation learning as a result of the limited size of the corpus or of sparse annotations, these methods cannot adapt to such situations. Another difficulty is the matter of determining how to quantify the adjustment when applying the correlation. Though concept similarity [9], sigmoid function [28], mutual information [10], random walk [24,25], random field [5], etc. have all been explored, this is still a challenge in the refinement of concept detections. In a state-of-the-art refinement method for indexing TV news video [8,6], the concept graph is learned from the training set. Though adaptation is considered to handle changes between training and test data, the migration of concept alinement to testing sets also depends on the affinity of two data sets, which is not always the case and can reduce the performance of indexing user-generated media, for example. Moreover, incomplete or imprecise annotations on training sets will further degrade the performance of these methods which rely highly on inter-concept correlations learned from training labels. The proposed TRF method in this paper is indeed a refinement methods but tries to tackle the above challenges.

These approaches to improving concept detection all try to compensate for the fact that it is really difficult to get accurate training data, i.e. annotations. TRECVid, the largest collaborative benchmarking activity in the area, with its collaborative annotation of training data among participants in one year realised a total of 8,158,517 annotations made directly by the participants of TRECVid or by the annotators of the Quaero project and a total of 28,864,844 annotations was obtained by propagating the initial annotations using the *implies* or *excludes* relations among concepts. While this may appear substantial and used clever techniques like an active learning procedure to prioritise annotations of the most useful sample shots [3] and to ask for a “second opinion” when manual annotations strongly disagreed with a