Applied Soft Computing 69 (2018) 453–463

Contents lists available at ScienceDirect

Applied Soft Computing

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

Interactive object retrieval via cost-minimizing queries

Weining Wu

College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, PR China

ARTICLE INFO

Article history:
Received 17 August 2017
Received in revised form 8 March 2018
Accepted 21 April 2018
Available online 5 May 2018

Keywords:
Interactive retrieval
Relevance feedback
Crowds learning

ABSTRACT

Studies using machine learning algorithms have documented advantage in applications of visual object retrieval. While interactive techniques have also been reported, the relevance feedback by minimizing the costs of time and annotation has not been investigated due to the requirement of data storage and online toolbox. In this work, we present an interactive scheme of object retrieval coupled with cost-minimizing queries in order to provide accurate results in a short time. We investigate a fast mechanism of relevance feedback to reduce time expense and explore the most reliable annotator in crowds to control annotation costs. Our experimental results demonstrate the effectiveness and efficiency of proposed framework.

1. Introduction

Given a visual concept described by some images, the goal of object retrieval is to find other images which may link to the same concept in the database [1,2]. In most existing approaches, it is formulated as a binary classification problem that a learner is firstly trained on a group of labeled examples, and then used to discriminate whether the new-gained examples belong to the same class. In the scenario of passive classification, a batch of training examples are randomly collected at once and annotated by precisely semantic labels. Hence, its faced problem is that the training set may include too many negative examples but excessively few positive examples.

For instance, there are normally hundreds of categories in the entire database and the number of images per category is tremendously different. For a special category, only a small fraction of images actually contain the target but the majority of candidate images are confirmed as different backgrounds. In order to adapt an effective classifier for such a task of unbalanced distribution between vast backgrounds and rare objects, a large number of training examples should be randomly annotated to ensure extensive coverage and sufficient representation for training a passive learner with high accuracy. Consequently, limited human resources are wasted in providing labels for training examples.

In order to overcome the distribution bias, the system of interactive retrieval [3,4] is designed as a more favorable framework for training a learner than passive learning. By borrowing active sampling strategy, interactive retrieval systems are able to sequentially select the most uncertain image of all as relevance feedback for label acquisition. Then, the system constructing on a few but helpful images can have similar or even better performance compared with that on lots of randomly labeled images. Therefore, annotation costs have been reduced by a large margin with the advent of interactive techniques [5,6]. Even though lots of research focusing on cutting down labeling burden of retrieval systems has been reported until now, its computation costs and annotation costs still remain as an important issue in most real tasks. Considering the two scenarios as follows.

Most existing systems select the image locating close to hyperplane boundary as the most confused one of all for relevance feedback. This close-to-boundary strategy has strong theoretical basis that the selected point can half the hyperplane space [7] and it has been proved effectively in practice [8]. But time expense in computing the distances between individual images and the hyperplane is overlooked in most existing works. When the retrieval system is employed on the large-scale data, its time for feedback should be considered, even the simplest function for calculating these distance is used. In practice, it needs to design an efficient feedback strategy for interactive retrieval systems.

On the other hand, existing retrieval systems assume that all labels for training data are generated from human experts, which makes it exceedingly expensive in the label acquisition and needs a significant amount of time. Meanwhile, the online toolbox [9] enables retrieval systems to collect a group of annotations at low costs from multiple volunteers manually in a short time, but it is hard to distinguish reliable annotations because the volunteers may come from a diverse pool including luminaries, experts and novices [10]. Consequently, it desires an effective method to generate reliable and low-cost labels in applications.

In this paper, we handle the above issues by proposing an interactive retrieval system via cost-minimizing queries. By extending the traditional scheme, we develop a fast strategy of relevance feed-

E-mail address: wuweining@hrbeu.edu.cn

https://doi.org/10.1016/j.asoc.2018.04.042
1568-4946/

back to reduce time expense, and then obtain reliable labels from online annotations to control labeling costs. Our contributions can be summarized as follows:

- We represent all images by compacted descriptors and train a sparse classifier as the learner in the retrieval system. Then, in every round, we compute the distances between candidate images and query hyperplane by requiring partial scores given by the current learner. Such distances can be computationally efficient, and then time expense in the feedback procedure is remarkably reduced.
- We use the technique of label estimation to obtain the reliable annotation from a group of volunteers with unknown levels of expertise. The estimated label is added for training the learner, instead of the actual label provided by the expert. Then, total annotation costs are controlled in the querying procedure.

The paper is organized as follows: in Section 2, some related works are summarized; in Section 3, the interactive retrieval system via cost-minimizing queries is presented in detail; in Section 4, experimental results are given, at last, conclusions and discussions are shown in Section 5.

2. Related works

In practice, it is an extremely hard procedure to collect sufficient images with precise annotations for training a learner in most retrieval tasks, especially when there are hundreds of categories in the database. Constructing a training set by randomly labeling images has been considered as a naive method in the passive framework, because it usually returns a majority of negative examples that do not match to the target concept. Then, it needs extra labeling costs to obtain enough positive examples for an effective training set. In hopes of making full use of limited human resources, interactive retrieval systems are being investigated in many laboratories around the world, and a vast body of literature has been accrued in this regard using current techniques of active learning [11–14].

In a typical interactive retrieval system, a small number of the target and backgrounds are labeled as the training set. Then, a classification model is initialized for the interactive retrieval system. In every active round, the learner identifies the most confused one from remained images, returns the selected image to human experts for precise annotation and adds the example-label pair for updating the model. After some rounds of feedback, the performance of retrieval system can be iteratively enhanced. Due to its nature in the training procedure, the interactive retrieval system usually concludes two main elements, i.e. return the confused image and acquire its annotation. Hence, we summarize the existing works as follows.

The identification of the confused image relies principally on data storage. In early studies, collected images are directly stored as a group of vectors by extracting their low-level feature. Then, the image which can minimize the distances between individual images and query hyperplane is considered as the most confused one of all [7,8]. Although it has been demonstrated as a simple and intuitive strategy for relevance feedback, recent works show that its time expense increases with the size of the database, because the learner needs to scan all vectors in the entire database. Recently, some researchers project the features of collected images into sparse histogram of quantized feature [15–17] or use the predictions of some binary classifiers pre-defined in the visual phrases [18,19], then measure the similarity in forms of approximate weighted histogram intersection [20] or approximate near neighbor methods [21,22]. These works show their efficiency in the tasks of point-to-point search, but they can not adapt to point-to-hyperplane search.

Furthermore, the acquisition of precise label for the returned image has also been studied [10,23]. In the universe paradigm of interactive retrieval systems, the “ground-truth” is provided by querying human experts, which makes it expensive and time-consuming in the annotation procedure [24]. In order to overcome the financial burden, the labels estimated from multiple online annotations have been used instead of the “ground-truth” from human experts. One common approach is to query all volunteers independently, then uses their voting results as approximate labels [25]. The other approach is to query reliable volunteers, then compute actual labels according to their levels of expertise [24]. In applications, the latter method can obtain more accurate labels, but needs less labeling expense than the former one, because not all volunteers are queried but only the reliable ones.

Generally, the total costs in constructing an interactive retrieval system include two aspects: time expense in relevance feedback and labeling expense in querying the annotator. In this work, we build such a retrieval system by cost-minimizing queries from these two aspects, i.e. the time expense is reduced by proposing a fast mechanism of relevance feedback with efficient computation, and the labeling expense is controlled by querying a reliable annotator instead of asking a human expert or multiple annotators.

3. Method

We begin by defining some basic notions and necessary preliminaries. Let $X = \{x_i\}_{i=1}^{n}$ be training examples and $Y = \{y_i\}_{i=1}^{n}$ be corresponding labels. For every observed image, we use the predictions from d binary classifiers as its feature and store it as a binary vector, i.e. $x \in X$, $x \in \{0, 1\}^d$. For a special category, in the 1-vs-all scheme, we have $y \in Y$, $y \in \{0, 1\}$. Then, a model $f_\theta : X \to Y$ can be learned on $D = \{(x_i, y_i)\}_{i=1}^{n}$ and the y can be estimated by the conditional probability $p(y|x, \theta)$ with the parameter θ. In the interactive framework, there is also a pool of unlabeled data $U = \{x_i\}_{i=n+1}^{n+m}$, $m \gg n$. In every round of interactive framework, the most confused image of all is returned as follows

$$x^* = \underset{x \in U}{arg\,max} \left(1 - \max_{y \in Y} p\left(y|x, \theta\right)\right) \tag{1}$$

In order to reduce annotation expense in querying human experts, in this work, we use online annotations coming from multiple volunteers instead. Similarly, suppose that there are L volunteers, then their annotations for x_i can be represented as $z_i = \{z_i^l\}_{l=1}^{L}$, $Z = \{z_i\}_{i=1}^{n}$, $z \in Z$. Based on the observed data, the label of returned image x^* can be estimated as

$$y^* = \underset{y \in \{0,1\}}{arg\,max} p\left(y|x^*, z\right) \tag{2}$$

3.1. Fast relevance feedback

In this work, we train a sparse model as the classifier for the retrieval task, and then we attempt to choose the confused image as relevance feedback by minimizing point-to-hyperplane distance. Assuming that the hyperplane query passes the origin, i.e. the θ and the x are unit norm, the Eq. ((1)) can be written as

$$x^* = \underset{x \in U}{arg\,min} \left|\theta \cdot x\right| \tag{3}$$

Since the θ and the x both have d components, we have

$$x^* = \underset{x \in U}{arg\,min} \left|\sum_{j=1}^{d} \theta^j \cdot x^j\right| \tag{4}$$