Pattern Recognition Letters 98 (2017) 110–116

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Query specific re-ranking for improved cross-modal retrieval

Devraj Mandal, Soma Biswas*

Department of Electrical Engineering, Indian Institute of Science, Bangalore, Karnataka, 560012, India

ARTICLE INFO

Article history:
Received 2 March 2017
Available online 6 September 2017

MSC:
41A05
41A10
65D05
65D17

Keywords:
Cross-modal
Image-text
Re-rank

ABSTRACT

Cross-modal retrieval tasks like image-to-text, audio-to-image retrieval, etc. are an important area of research. Different algorithms have been developed to address these tasks. In this work, we propose a novel query specific re-ranking based approach to improve the retrieval performance of any given baseline approach. For each query, the top K-retrieved results of the baseline algorithm are used to compute its class-rank order feature. Based on this feature of the query and the highly relevant examples within the top K-retrieved results, each training example is given a score indicating its relevance to the query, which is finally used to train the query-specific regressor. The new score given by this regressor to each retrieved example is then used to re-rank them. The proposed approach does not require knowledge of the baseline algorithm, and also does not extract additional features from the data. Thus it can be used as an add-on to any existing algorithm for improved retrieval performance. Experiments with several state-of-the-art cross-modal algorithms across different datasets show the effectiveness of the proposed re-ranking algorithm.

1. Introduction

Cross-modal retrieval is an important area of research in the field of computer vision and pattern recognition with a wide range of applications. For example, given a text query, we may want to retrieve semantically meaningful images from the database. A few examples of cross-modal data matching considered in this work are shown in Fig. 1. Several approaches have been proposed in the literature to address this task [1–3].

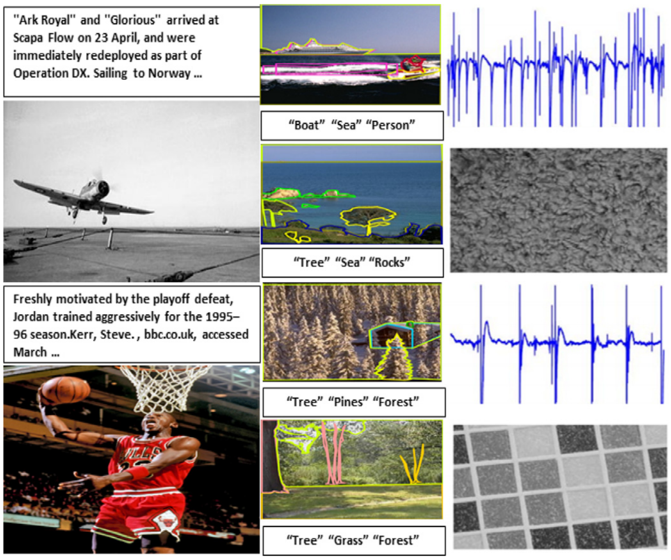
In this work, we propose a novel query specific re-ranking approach for improving the retrieval results of any baseline algorithm. The input to the algorithm is the retrieval results of the baseline approach along with their similarity (or distance) scores and also the training data used by the baseline approach. The majority of the cross-modal approaches aim to find the relation between the different modalities. In this work, we ask the question: *can we use the relative positioning of the query and retrieved data with respect to its own modality to improve the retrieved results?* For each query, first a class-rank order feature is computed based on the top K-retrieved results of the baseline algorithm. Based on this new feature, a subset of the top K-retrieved results, termed as the highly relevant set is chosen. This is based on our confidence as to which of them actually belong to the same class as the query. These are used to compute a score for each training example, which is then used to train the query-specific regressor. The new score given by this regressor to each retrieved data is finally used to re-rank them.

Extensive experimental evaluation is performed on several baseline algorithms and on different datasets to justify the usefulness of the proposed re-ranking approach. Specifically, for the baseline algorithms, we use a variety of classical approaches like Canonical Correlation Analysis (CCA) [1,2], dictionary learning approaches like Coupled Dictionary Learning (CDL) [4], deep learning based approaches like Deep CCA with Auto-encoders (DCCA-AE) [5] and so on. The approach has been tested on a variety of multi-modal databases like Wikipedia [6], LabelMe [7], Materials [8] and Multiple Features [9] and significant improvements over the Mean Average Precision (MAP) and rank-1 accuracy has been observed. The contributions of this work are as follows:

1. We propose a novel query-specific re-ranking framework, which is able to improve the baseline retrieval results for cross-modal tasks and can be used as an add-on to any existing approach.
2. The proposed approach does not generate new features, and also does not require knowledge of the inner workings of the baseline method.
3. The approach is able to improve the baseline results of various approaches over a variety of different datasets involving a wide variety of cross-modal retrieval tasks.

* Corresponding author.
E-mail addresses: devraj89@ee.iisc.ernet.in (D. Mandal), soma.biswas@ee.iisc.ernet.in (S. Biswas).

http://dx.doi.org/10.1016/j.patrec.2017.09.008
0167-8655/© 2017 Elsevier B.V. All rights reserved.



"Ark Royal" and "Glorious" arrived at Scapa Flow on 23 April, and were immediately redeployed as part of Operation DX. Sailing to Norway ...

Freshly motivated by the playoff defeat, Jordan trained aggressively for the 1995–96 season.Kerr, Steve., bbc.co.uk, accessed March ...

Fig. 1. A few examples of cross-modal data matching. First two columns: image-text modalities (Wikipedia and LabelMe datasets); Third column: audio signatures and corresponding images of texture surfaces (Materials dataset).

The rest of the paper is organized as follows. Section 2 discusses the related works. Details of the proposed approach are described in Section 3. The experimental results are given in Section 4 and the paper concludes with a brief discussion.

2. Related works

Here we discuss some of the related works in the literature on cross-modal matching as well as re-ranking.

Cross-modal approaches: First, we discuss the relevant literature for cross-modal matching. Given paired data of two modalities, Canonical Correlation Analysis (CCA) [1,2] learns a lower dimensional feature space from the two modalities. To handle the non-linear relationship between the data, kernel trick has been employed to devise Kernel CCA [1,2]. The constraint of paired data has been removed in the formulation of mean CCA, cluster CCA [10] and their kernelized versions. Generalized Multiview Analysis (GMA) [3] mathematically formulates the cross-modal analysis problem as a constrained quadratic program and provides a solution by generalized eigenvalue approach. GMA is shown to be a supervised extension of CCA and an extension towards its kernel form has also been designed. Scalable variants of CCA has been developed by using randomness to capture the intrinsic non-linear relationship between data from two modalities [11].

As an alternative to the non-parametric KCCA, Deep Canonical Correlation Analysis (DCCA) [12] uses neural networks to learn complex non-linear transformations of the two views of data such that the resulting representations are highly linearly correlated. Auto-encoders combined with a deeper representation of CCA in a jointly optimized framework has also been devised for cross-modal retrieval tasks [5]. Relatively Paired Space Analysis (RPSA) [13] uses the relative pairing of information to build a discriminative latent model while solving a maximum margin problem. Coupled Dictionary based Learning Methods (CDL) [4] learns two projections over the sparse representation domain to learn a common subspace for cross-modal matching.

Re-ranking approaches: Now, we will discuss the relevant literature on re-ranking. In [14], a novel clustering algorithm for tagging a face dataset has been provided. The rank-order distance is motivated by the observation that faces of the same person usually share their top neighbors. The rank-order distance has been extended for solving person re-identification problems [15]. This problem has also been solved by using soft biometric attributes [16], appearance attribute subspaces [17], learning intra-camera discriminative models [18], and even bi-directional ranking methods [19] involving both the probe and gallery data as the query. Deep metric learning [20] involving a siamese deep neural network is used to learn the color feature, texture feature and a metric jointly for practical person re-identification. Visual saliency and consistency have been used to design a re-ranking algorithm in [21]. Contextual spaces aiming to explore relationships between images have also been used for re-ranking [22].

A few more re-ranking approaches proposed in the literature can be found in [23–26]. These query based adaptive re-ranking methods usually generate positive and negative image pairs for each query to re-rank the results. Though these methods show impressive performance in the person re-identification problem, they are mainly suitable for image-image matching and it is not clear how they can be extended for the cross-modal scenario. The POP algorithm in [24] uses a user-based interactive "one-shot" approach to identify positive and negative image examples to improve re-identification. The work in [26] proposes an iterative approach in a dictionary learning framework to handle re-ranking problems. In contrast, the proposed framework works with the original features and thus can be used with any existing algorithm provided the training data and the original distance scores are available. The re-ranking concept is slightly different from fine-tuning a deep model, where the model is usually first trained on a large auxiliary dataset and then fine-tuned to adapt to the target dataset. For this work, there is just one training dataset, and the proposed algorithm utilizes the same features that is used by the baseline algorithm.

Now, we will briefly describe some popular regression techniques which are used in our work [27]. Regression is a statistical process for estimating the relationships between a dependent variable and one or more independent variables. Linear regression [27] expresses the dependent variable using linear functions where the model parameters are estimated by studying the provided sample data. Linear regression (LR) has a simple closed form solution with techniques to use different regularization constraints to prevent over-fitting. Support Vector Regression (SVR) [27] finds a function which is bounded by a particular deviation from the target variables. In addition, SVR gives the option to embed the input data into high dimension using kernels and hence exploits the concept of non-linearity to achieve better performance in general.

3. Proposed approach

In this section, we describe in detail the proposed query-specific re-ranking algorithm.

3.1. Problem definition

Let the two modalities be denoted by **X** and **Y** and let the training data for the two modalities used by the baseline algorithm be denoted by $\mathbf{X}_{tr}$ and $\mathbf{Y}_{tr}$ respectively. Let the labels of the training data be denoted as $\mathbf{X}_{tr}^{L}$ and $\mathbf{Y}_{tr}^{L}$, where $L = \{1, 2, \ldots, C\}$, with C being the total number of classes. For a given query X_q and a baseline algorithm, let $\mathbf{Y}_q = \{Y_1, Y_2, \ldots, Y_N\}$ be the retrieved results with distances $\{d_1, d_2, \ldots, d_N\}$ from the query, where, $d_{i+1} \geq d_i$, for $i = \{1, 2, \ldots, N\}$. Given this information, the goal is to re-rank the retrieved data $Y_1, Y_2, \ldots, Y_N$ for improved retrieval performance as compared to the baseline algorithm, without any other information regarding the actual principles of the baseline algorithm.