Neurocomputing 149 (2015) 1658–1671

Contents lists available at ScienceDirect

Neurocomputing

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

Image annotation based on feature fusion and semantic similarity

Xiaochun Zhang[a,*], Chuancai Liu[b]

[a] School of Information Engineering, Anhui University of Finance and Economics, Bengbu, 233030, China
[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

ARTICLE INFO

Article history:
Received 8 March 2013
Received in revised form
26 June 2014
Accepted 13 August 2014
Communicated by Xiaoqin Zhang
Available online 21 August 2014

Keywords:
Feature fusion
Linear transform
Nearest neighbor
Semantic similarity
Canonical correlation analysis

ABSTRACT

The present study developed a new algorithm based on multi-feature fusion and semantic similarity for image annotation. To study the relationship between feature distance and semantic similarity, the feature-annotation space is transformed into a distance–similarity space. Therefore, the intrinsic statistical relationship between visual and semantic information can be studied. Re-scaling is necessary to fuse multiple features. The potential multimodal properties of image features mean that traditional feature re-scaling is based on boundary values, which are sensitive to outliers. Our proposed distance re-scaling method overcomes the drawbacks of using statistical information. In the distance space, each distance vector of an image pair is treated as a sample. The anisotropic Gaussian distribution is transformed into an isotropic Gaussian with a mean of zero and standard variance. The nearest-distance images are retrieved from this space. To select features of not only low distance correlations but also a high semantic correlation, the visual and semantic relationship is studied using canonical correlation analysis. The canonical correlation coefficient of the similarity and distance is found to connect closely with the annotation score of the feature. Experiments showed that the proposed multi-feature fusion method removed the effects of scale and the correlations of feature distances, so it could represent the total distance better and find the nearest neighbors. We tested our method using the Corel5K, IAPR-TC12, ESP Game, and VOC PASCAL datasets, which showed that it outperformed existing approaches.

1. Introduction

Image annotation is an active research area in the computer vision domain. Most existing image annotation approaches can be classified into three categories: probabilistic methods; classification methods and nearest neighbor-based methods.

Probabilistic methods have been used successfully for image understanding [1–4]. Probabilistic topic models are a suite of algorithms, which aim to discover hidden data structures in large archives of discrete data [5]. Latent topic model researchers aim to determine the relationships between different types of data. These models learn the joint probability distribution between image features and text annotations [6–8]. Image understanding based on semantic information still has problems solving the semantic gap problem between high-level semantics and low-level visual features [9,10]. However, solving the semantic gap is still a tough problem. In the present study, we tried to determine the relationships between visual and semantic features using experiments.

Discriminative models can be regarded as a type of multi-class image classification with a very large number of classes [11–14]. These models train classifiers for every label and predict the class label for the test image, before copying the tags from the labels of images in the predicted class. A hybrid algorithm of a discriminative scheme and the nearest neighbor method use the neighborhoods of test images [15]. A new nearest neighbor model was proposed in [16], which used a weighted combination criterion to predict tags. Some online annotation platforms are also emerging in the commercial world [17,18].

The nearest neighbor approach is a naïve but efficient technique, which has received much attention for automatic image annotation. Recently, Makadia et al. proposed a simple *K*-nearest neighbor method, which was comparable to most state-of-the-art algorithms in terms of performance [19]. The total distance is computed from several types of normalized features, which is known as the "Joint Equal Contribution" (JEC). Zhang et al. described a similar approach with a partial implementation code [20], which was a nearest neighbor method because most of the annotation words were derived from the nearest image. This method extracted multiple features and computed the total distance. However, we found that multiple features might not significantly outperform a single feature. The present study was focused on this observation.

Traditional semantic similarity is a concept where a set of documents or terms in term lists are assigned to a metric based on

[*] Corresponding author.
E-mail address: zhang_xc@qq.com (X. Zhang).

http://dx.doi.org/10.1016/j.neucom.2014.08.027
0925-2312/© 2014 Elsevier B.V. All rights reserved.

their similar meanings. For computer vision application, the semantics and its similarity have no generally accepted definition. To study semantics quantitatively, we treated the textual annotation as the semantics and we defined a function to measure their similarity.

We studied the relationship between the semantic space and feature space, and we developed a feature fusion scheme based on statistical distance information. This new scheme may be a better candidate than the original, which is based on the boundary values of features. The main contributions of this study are (1) we defined a metric to measure the semantic similarity, (2) we investigated the visual–semantic relationships in the distance–similarity space and (3) we developed a distance normalization and multi-feature fusion method.

The rest of this paper is organized as follows. The proposed framework and algorithm are presented in Section 2. Section 3 presents the experiment and our evaluation. Finally, we conclude our work in Section 4.

2. Visual–semantic space

The task of automatic image annotation involves the prediction of textual annotations for a test image. The annotations may comprise one or several words. The predictions can be obtained from training images and related annotations. However, the raw images and annotations entail several problems and a suitable representation is required before learning.

Images and annotations can be transformed into a feature-annotation space and each image and corresponding annotation will become a point in this space. Several types of features may coexist. One way of retrieving annotations from test images is to find the most visually similar training images and using their annotations.

2.1. The problem of the original approach

It is helpful to review the original approach and its problem. From a set of images, multiple features (e.g. u, w) are extracted and represented as matrixes. Each column vector of the matrixes corresponds to an extracted image feature. Eq. (1) defines the matrix and vector notation used in this study where the dots denote all of the elements in the index:

$$a=\begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix}, \quad a_{i,}=(a_{i,1}\cdots a_{i,n}), \quad a_{,j}=\begin{pmatrix} a_{1,j} \\ \vdots \\ a_{m,j} \end{pmatrix} \tag{1}$$

The distance between features measures the visual similarity. For example, a color histogram feature with a short distance indicates that the images share a similar color appearance.

To compute the total distance between multiple features for images c and d (Eq. (3)), normalization is performed using the boundary values (Eq. (2)). Jain et al. summarized some normalization approaches for multimodal distributions [21], such as z-score, tan h, etc. These normalizations can be applied to image features for their multimodal nature. The z-score may be an appropriate option for the features resembling Gaussian distribution. A following experiment records the annotation performance of some normalizations:

$$\tilde{u}_{i,j}=\frac{u_{i,j}}{\sum_k(\max(u_{k,})-\min(u_{k,}))} \tag{2}$$

The value of "dist" with a superscript is a suitable distance metric (e.g., L_1 or L_2) for a feature type. The combination of Eqs. (2) and (3) is referred to as JEC [19]. Other methods such as lasso, group lasso, least square, and L_2 regularization also use Eq. (2) to normalize the features before computing the relevant total distance [19,20]:

$$\begin{aligned} distance(c,\ d) &= dist^u(c,\ d)+\ \cdots\ +dist^w(c,\ d) \\ &= \left|\tilde{u}_{,c}-\tilde{u}_{,d}\right|+\cdots+\left|\tilde{w}_{,c}-\tilde{w}_{,d}\right| \end{aligned} \tag{3}$$

It is obvious that the elements of the features may follow any distribution, such as uniform, Gaussian, or multinomial distributions. The boundary values are not reliable for normalizing the features. An extreme example is provided, as follows. Suppose that for one image, all of the elements in feature u are 100, whereas for other images, the values are between 0 and 1. Incorrect normalization means that feature u will contribute almost nothing to the total distance. The nearest neighbors are not necessarily the most visually similar images so the annotation retrieved may not be correct. Thus, to eliminate the effects of outliers, normalization should be based on statistical information (i.e., the mean and covariance).

The implementation described by Zhang [20] used L_1-norm for all of the features. For simplicity, this convention was used in the present study with a minor exception. The L_2-norm was used for Gabor feature, which agrees with other previous implementations [22]. The L_1-norm is also used for other features, including HOG and color histogram features. In addition, choosing different norms allowed us to investigate whether this discussion is applicable to other distance norms.

2.2. Space transform

Fig. 1 shows the components that comprise the space transformation. The first step is to extract the annotation tags and features from a pair of training images. The second step is to compute the distance between the features and the similarity between the tag vectors.

The semantic similarity between two images, c and d, is computed using Eq. (4). Eq. (5) shows an example with the keywords "person", "street", and "college". The similarity between

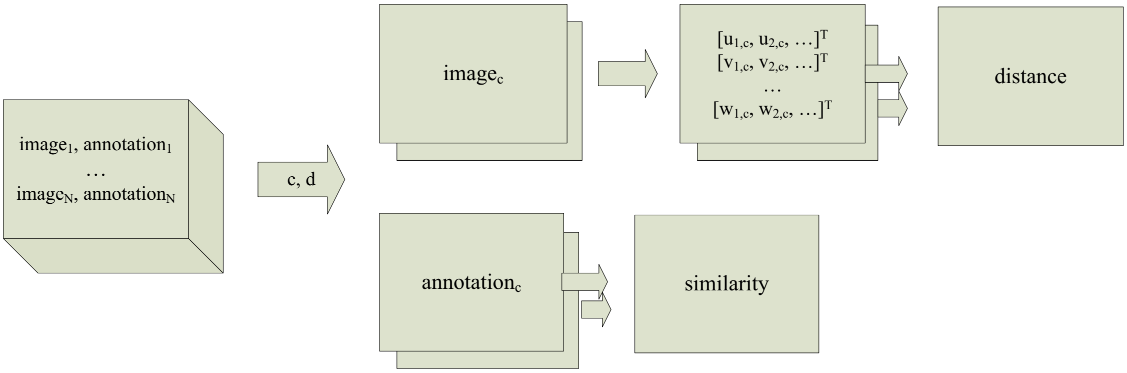
image$_1$, annotation$_1$
…
image$_N$, annotation$_N$
c, d
image$_c$
[u$_{1,c}$, u$_{2,c}$, …]^T
[v$_{1,c}$, v$_{2,c}$, …]^T
…
[w$_{1,c}$, w$_{2,c}$, …]^T
distance
annotation$_c$
similarity

Fig. 1. Visual–semantic space transformation.