Automatic Image Annotation using PHOW Features

Shereen A. Hussein, Howida Youssry Abd El Naby, Aliaa A. A. Youssif

Abstract

Image annotation is a task of assigning a set of semantic tags or keywords to unlabeled image based on a training set of data. This machine learning process depends on extraction and clustering the low-level features of images then mapping them to the semantic which is high-interest image retrieval. Many annotation techniques with reasonable performance have proposed in the last decade. The proposed algorithm for automatic annotation in this paper depends on similarity computation and label transferring to the query image. Similarity computing uses low-level image features and significant distance estimates. The feature vector of the test image compared with the feature matrix of similar and dissimilar image pairs in the training data sets. Then the label or keyword transfer from the similar image pair to the test image is performed by counting the local frequency of neighbor’s keywords. Performance is evaluated using precision and recall.

Keywords: Automatic Image Annotation, Similarity, Feature Matrix, PHOW.

1. Introduction

Automatic image annotation is an interesting topic in current searches due to its effective impact in image retrieval systems. It concerns in labeling the digital images with keywords express their contents. As a machine learning process, [1-4] the correlation between extracted image features and the training annotation word list are used translate the textual vocabulary to visual vocabulary. The huge challenge in this process is to connect the various types of low-level features (colors, textures, and boundaries) to a high-level concept.

Automatic image annotation is regarded as a kind of multiclass image classification [5] and also known as linguistic indexing or automatic image tagging. Converting automatic image annotation into classification problem, make the annotation process [6-8] easy under a standard semantic label. Therefore, feature extraction and selection is a vital strategy to guarantee reliable and meaningful results for data classification alongside different advantages such less data storage and computation cost. First, the extracted features are modeled in pyramid histogram of visual words then converted to form feature matrix then compute the similarity between image pairs to facilitate query image automatically annotation. The rest of the paper organized as follows. Section 2 gives a short review of automatic image annotation. Section 3 presents the proposed annotation method using PHOW features and similarity image pairs; Section 4 shows the experimental results and analysis of the annotation on Corel5k and IAPR TC-12 dataset. Section 5 contains the conclusions and future work.

2. Related Work

Mori et al. [9] proposed the first attempts at image auto-annotation as transforming image to word based on word co-occurrence. The word co-occurrence is a linguistic term used in the natural image processing which refers to the often used words together. Duygulu et al. [10] proposed the image auto-annotation in a machine translation model. Focusing on region-based image annotation instead of the global image annotation because of its insufficiency in determine which part in image relate to which label. Jeon et al. [11] introduced the Cross-Media Relevance Model (CMRM) where visual information of an image denoted as blob set. It helps in clarification the image semantic information obviously. Putthividhya D. et al. [12] used Probabilistic latent semantic analysis (PLSA) model and latent Dirichlet allocation (LDA) to uncover the hidden themes in documents collections. Feng S et al. [14] proposed Bernoulli relevance model to improve CMRM and CRM accuracy. It is a generative statistical model that uses an annotated training set for the annotation process. Tianxia Gong, Shimiao Li, Chew Lim Tan et al. [15] used probabilistic models to present a framework to represent the word-to-word relation. Test images prediction tags depends on the classifier of visual features that is trained by the discriminative model. S. Deerwester et al. [16] proposed Latent Semantic Indexing (LSI) which deals with the problems of referring the same object with multiple words or some word with multiple meanings. So it overcomes the traditional lexical matching techniques shortcomings. Hofmann et al. [17] used the probabilistic LSI (PLSI) model, as an alternative to LSI. The PLSI is a more
robust technique for automatic document indexing where each document represented by its word frequency. S. Zhang et al., [18] proposed group sparsity as a technique for automatic image annotation. It solves annotation as a retrieval problem and handles the features selection issue. Images pair similarity is evaluated by a specific value if it is positive means pair are similar otherwise are dissimilar. The overall performance is highly affected by pair similarity computations [19–28].

3. Proposed Framework
An efficient method for automatic image annotation is to adopt labels from similar images in the dataset with keywords. Doing so has two sub-problems first, computing the image similarity and second, choosing the keywords from the similar images. Simple methods for these sub-problems proposed in the framework of creating feature matrix of n feature vector for each image in the training set and pairs of similar & dissimilar images. Image similarity measure has built from analysis on the histograms of a bag of visual words features matrix. The similarity within the image pair considers the similarity of keywords. Positive image pairs share some keywords, while negative ones do not. Based on the weights calculated the most similar images are found out from pairs. The quality of these pairs highly influences the overall performance. However, the ground truth of similarity is not available.

3.1 Feature Extraction (The PHOW Descriptors)
A Feature matrix for the images dataset created with the PHOW feature vectors. Image description by Pyramid of Histograms of Visual Words (PHOW) method is an extension to the bag-of-words (BOW) model in which the extracted SIFT image features treated as words. It considers the local information feature of the image [29-32]. The method implemented as in figure 2. PHOW [33, 34] overcomes the drawback of BOW of unavailability of spatial image features information by dividing the image into fine sub-regions (pyramids) and concatenating the histogram of each of these regions to the histogram of the original image with a suitable weight. When color images were processed, they are converted from RGB space to HSV color space with the SIFT feature extracted from each channel. For grayscale images, only the intensity is used. Hence, the resulting SIFT feature dimension is 128x3 for color images and 128 for grayscale images. Once the SIFT features obtained, “bag of words” model is used to quantize them into visual words by k-means clustering. Thus the image is represented by a histogram of visual word occurrences.

![Figure 1 Proposed Framework Architecture](image1.png)

**Figure 1 Proposed Framework Architecture**

3.2 Obtaining Image Pairs
In this framework, the objective function is a similarity function [39-43]. Similar and dissimilar image pairs are required for training purposes. Since the goal is to assign relevant keywords, A. Makadia [7] proposed an approach to discover these pairs based on the keyword similarity [44-46]. Any pair of images in the training set that shares more than four keywords is considered as a positive training example, while a pair without any common keyword is a negative one. Obviously, the distances within positive pairs are expected to be smaller than the ones within negative pairs, since images of similar pairs should be much closer to each other than dissimilar ones as shown in figure 3.

![Figure 2 the PHOW Feature Matrix Steps](image2.png)

**Figure 2 the PHOW Feature Matrix Steps**
3.3 Weight Vector Calculation

The weight of each feature vectors calculated by using feature matrix obtained using the PHOW method and set of pairs of similar and dissimilar images. The most efficient method to compute the weight vector is weighted least square (1) due to its ability to provide different types of easily statistical interpretation for estimation and prediction.

\[ W = \arg \min || Xw - Y ||_2^2 \quad w \in \mathbb{R}^p \]  

(1)

In the image annotation testing stage, this weight is used for computing similarity array by the iterative equation (2) for each row in the feature matrix. 

\[ S_i = (t_i - f_i) \times w_i \]  

(2)

Where: 
- \( t \) is feature vector of test image. 
- \( f \) is feature value of feature matrix of training images. 
- \( w \) is the weight vector.

Then a comparison between the similarity array and the pairs of similar and dissimilar images is performed to choose a predefined number of values from \( s \) with the highest local frequency keywords which are transferred to annotate the image.

4 Experiments & Discussion

Here is a description of used benchmarks (Corel5K, IAPR TC-12) and evaluation of image annotation method.

4.1 Dataset

In Figure 4, there is some samples from Corel5K and IAPR datasets.

- Corel5K [47, 48] has become a de-facto evaluation benchmark in the image annotation community. It contains 5,000 images collected from the larger Corel CD set, split into 4,500 training and 500 test examples. Each image is annotated with an average of 3.5 keywords, and the dictionary contains 260 words that appear in both the train and the test set.

- IAPR TC-12 [49] is a collection of 19,805 images of natural scenes that include different sports and actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life. Unlike other similar databases, images in IAPR TC-12 are accompanied by free-flowing text captions. While this set typically used for cross-language retrieval, we have concentrated on the English captions and extracted keywords (nouns) using the Tree Tagger part-of-speech tagger. That resulted in a dictionary size of 291 and an average of 4.7 keywords per image. 17,825 images used for training and the remaining 1,980 for testing.

4.2 Discussion

Understanding the scene is the key to solving the image annotation problem at the human level. However; objects identification and giving them keywords based on the given scene is still a hot topic in research. The goal of this work was not to develop a new annotation method but enhance the performance with presenting multiple techniques of similarity computation between image features. Experiments [50-55] on the two different datasets Corel 5K and IAPR TC12 aim to bridge the gap between the low-level representations of images which concerned with color, shape and texture and the high-level one with the semantic meanings. It is clear that a simple combination of basic distance measures over some new feature extraction methods can effectively serve in the performance of image annotation methods.
4.3 Performance Evaluation

The performance of annotation task calculated by F-score value (5) which uses the value of precision (3) and recall (4). The proposed algorithm achieves F-score 0.45 for Corel5K dataset and 0.42 for IAPR TC12 dataset. Figure 5 shows an example of the annotation results in comparison to the human labeling and Table 1 presents a list of F-score values for different annotation models versus the proposed method.

$$\text{Precision} = \frac{\text{no. of correct annotated label}}{\text{Total annotated label}} \quad (3)$$

$$\text{Recall} = \frac{\text{no. of correct annotated label}}{\text{Total label in test set}} \quad (4)$$

$$F - \text{score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{(\text{precision} + \text{recall})} \quad (5)$$

Table 1 Annotation results of the proposed method versus the previous published algorithms on Corel5k & IAPR TC12 Datasets [ZHANG et al.][5]

<table>
<thead>
<tr>
<th>Model</th>
<th>F-Score Corel5K</th>
<th>F-Score IAPR TC12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-occurrence</td>
<td>0.02</td>
<td>-</td>
</tr>
<tr>
<td>Translation Model</td>
<td>0.05</td>
<td>-</td>
</tr>
<tr>
<td>CMRM</td>
<td>0.09</td>
<td>-</td>
</tr>
<tr>
<td>Max. Entropy</td>
<td>0.10</td>
<td>-</td>
</tr>
<tr>
<td>CRM</td>
<td>0.17</td>
<td>-</td>
</tr>
<tr>
<td>MBRM</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td>GS</td>
<td>0.31</td>
<td>0.30</td>
</tr>
<tr>
<td>Tag-Prop</td>
<td>0.36</td>
<td>0.39</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>0.45</td>
<td>0.42</td>
</tr>
</tbody>
</table>

5 Conclusion & Future Work

In the proposed framework, pyramid histogram of visual words description used in extracting image features to solve the automatic image annotation problem. Since the BOW image model for a feature, representation is the base for image annotation and classification. Weighting these features using weight least square method and obtaining accurate similar and dissimilar pairs of images is enhancing the power of assigning correct labels for the image that help in understanding its meaning. We tried in this algorithm to get used of PHOW advantages in formulating certain features. Also combining similarities computation between training image pairs with the keywords information gives high results for the annotation task in comparison to other existing methods. In future work, try using ontologies which are the integration of images analysis improvement knowledge and images interpretation, instead of using contextual knowledge for semantic image annotation by building semantic hierarchies.

6 References

International Conference on Computer Vision and Pattern Recognition, 2006.
[34] Bosch Anna, Zisserman Andrew and Munoz Xavier, "Representing shape with a spatial pyramid kernel," in CVIR, Amsterdam, 2007.