Neurocomputing 213 (2016) 115–124

Contents lists available at ScienceDirect

Neurocomputing

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

Part-based clothing image annotation by visual neighbor retrieval

Guang-Lu Sun, Xiao Wu *, Qiang Peng

School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China

ARTICLE INFO

Article history:
Received 19 September 2015
Received in revised form
9 December 2015
Accepted 9 December 2015
Available online 29 June 2016

Keywords:
Image annotation
Clothing search
Part-based
Annotation by search

ABSTRACT

With the advent and popularity of e-commerce and clothing image-sharing websites, clothing image search and annotation become active research topics in recent years. Clothing image annotation is a challenging task due to large variations in clothing appearance, human body pose and background. In this paper, we explore part-based clothing image annotation in a search and mining framework. Similar image search is first conducted to discover visual neighbors for a query image. The impact of large variations of clothing is alleviated by pose detection and part-based feature alignment. Tag relevance and tag saliency are taken into consideration to obtain the candidate tags. The relevance of candidate tags is identified by mining visual neighbors of a query image, while the saliency is determined according to the relationship between query image parts and part clusters on the whole training set. Experiments on a dataset with 1.1 million clothing images demonstrate the effectiveness and efficiency of the proposed approach.

1. Introduction

In recent years, the rapid development of social network and image-sharing websites leads to the explosion of web images, which requires effective image search techniques. Image annotation, aiming to describe image content with appropriate textual tags, has been an active topic in computer vision and machine learning areas. Nowadays, more and more people prefer online clothing shopping due to its convenience and attractiveness, which makes clothing shopping a huge market. It greatly motivates clothing relevant research [1–7]. However, there exist limited studies on clothing image annotation, which can directly help customers find their favorite clothing with representative attributes, such as “blue boat-neck dress”, “polka-dot lace skirt”. Therefore, part-based clothing image annotation is a valuable and meaningful task to explore.

Existing annotation researches for general images can be roughly classified into two categories: model-based [8–24] and annotation by search [25–36] approaches. Model-based methods are similar to multi-label classification. They usually train a set of tag classifiers by supervised learning and use their outputs to annotate images. Unfortunately, the number of tags for web images is unlimited and increasing constantly. It is infeasible to train a classifier for each tag. Different from model-based methods, visual-neighbor-based methods, also called annotation by search, mainly consist of two steps: (1) Visual neighbors are first retrieved by calculating the visual similarity between a query image and the whole dataset, and then associated tags of visual neighbors are treated as candidate tags. (2) Tags are re-ranked by estimating the relevance of each candidate tag to the query image in certain manner. Since visual-neighbor-based methods are unsupervised learning methods, which do not need to learn a mapping between low-level features and high-level semantic concepts, they are more scalable and tend to be more suitable for web image annotation due to their concision and effectiveness.

Clothing image annotation is a challenging task due to complex backgrounds, diverse appearances and high user expectation. The background noise of clothing images in natural scene is a considerable factor affecting visual search results. More importantly, user expectation for clothing image annotation is quite different from general image annotation. The latter is an objective-level representation. However, clothing image annotation is a detailed description for clothing local attributes. It is desired that the top-ranked tags of a clothing image are not only relevant to clothing image content, but also can reveal the most prominent features of the clothing. Fig. 1 illustrates two web images with tags. For the general image on the left, tags such as “sunset”, “sky”, “beach”, and “sea” are relevant to image content, which demonstrate the object-level description of this image. For the clothing image on the right, although general tags such as “coat” and “T-shirt” are relevant to image content, they are not enough to describe the representative features of this clothing. Users need more unique and specific

* Corresponding author.
E-mail addresses: sunguanglu66@126.com (G.-L. Sun), wuxiaohk@home.swjtu.edu.cn (X. Wu), qpeng@home.swjtu.edu.cn (Q. Peng).

http://dx.doi.org/10.1016/j.neucom.2015.12.141
0925-2312/

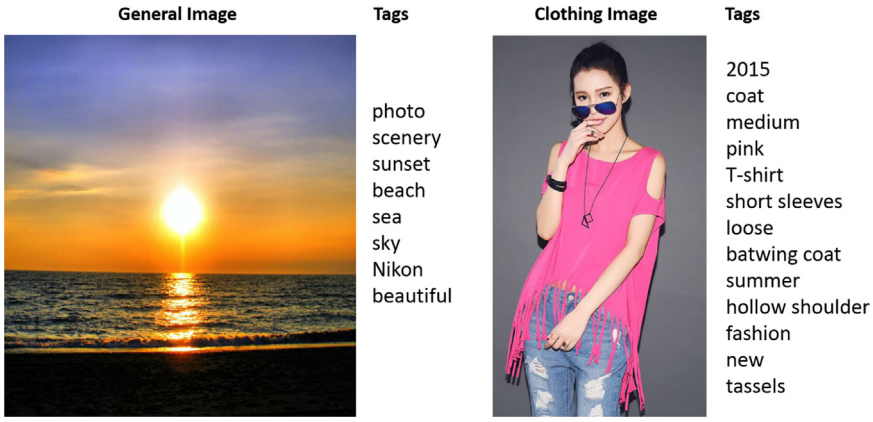

Fig. 1. Examples of two images with tags (a) general image and (b) clothing image.

details like "hollow shoulder", "tassels" to better describe its salient features. Therefore, two key issues should be addressed for clothing image annotation: how to effectively search visual neighbors for a clothing image and how to assign relevant and representative tags to it.

For traditional image annotation by search [34–36], the visual relevance of a tag to a query image is measured by calculating the visual similarity between the image and images with the same tag. However, for clothing annotation, it is obviously unreasonable. For clothing image annotation, users prefer tags for clothing details, which are usually the most prominent features of the clothing. The analysis of visual relevance and semantic relevance on whole images may not be suitable for clothing image annotation, especially for certain clothing attributes. For instance, tag "v-collar" is a description for clothing collar, which is only one part of the clothing. Its corresponding visual regions should be neck, rather than arms, shoulders, or other parts of the human body. This is not to use the visual features of the whole image for the relevance estimation. Therefore, for certain tags representing clothing local attributes, part-based relevance analysis between tags and images is more meaningful.

In this paper, we propose a part-based clothing image annotation approach under the framework of annotation by search, which takes into account tag relevance and tag saliency. With the assistance of pose estimation, clothing image search is conducted to obtain a robust performance by part-based feature alignment. Part-based salient tag extraction method is proposed to estimate the tag relevance, which is partial description of the clothing. Finally, it is fused into a unified annotation by search framework to obtain better clothing annotation results.

The main contributions of this work are summarized as follows:

- A part-based clothing image annotation approach under the framework of annotation by search is proposed, which takes into account tag relevance and tag saliency.
- A part-based salient tag extraction method is adopted to select dominant tags for clothing images, which combines relevance analysis of intra-cluster and inter-cluster based on part clusters.
- Experiments on a large-scale clothing image dataset with 1.1 million images demonstrate that the proposed approach outperforms the state-of-the-art methods on image annotation.

The rest of this paper is organized as follows: related work is reviewed in Section 2. The part-based clothing image annotation is elaborated in Section 3. Experiments and results are presented in Section 4. Finally, this paper is concluded with a summary.

2. Related work

2.1. Clothing research

With the explosive growth of clothing images, there exist a number of clothing-related studies, such as clothing image retrieval [1], clothing image segmentation [2,3], clothing recommendation [4] and clothing attribute learning [5–7]. A practical problem of cross-scenario clothing retrieval is addressed in [1] by combining pose estimation and transfer learning technology. Clothing extraction methods [2,3] obtain the clothing objects by distinguishing the foreground and background regions. To more accurately identify clothing regions, the foreground region is predicted with face and skin detection [2]. An automatic fashion image parsing with weak labels is addressed in [3], which combines human pose estimation, MRF-based color and category inference, and superpixel-level category classifier learning to parse fashion items. An occasion-oriented clothing recommendation system [4] is constructed by considering two key criterions: wear properly and wear aesthetically. Clothing attribute learning is essentially model-based works for clothing image annotation [5–7], which models the relationship between well-defined attributes and low-level features. In this paper, we attempt to carry out part-based clothing image annotation by visual search.

2.2. Image annotation

As mentioned formerly, model-based methods focus on modeling the relationship between image features and tags. A probabilistic formulation for semantic image annotation is proposed in [8], which shows that a minimum probability of error annotation is practical by establishing the correspondence between semantic labels and semantic classes. An ensemble of binary classifiers is trained to predict the label membership, and then image annotation with semantic labels is carried out to mine the label membership of images [9]. A probabilistic topic-connection model is built to represent the connection between text descriptions and images [10]. A relevance model is trained by analyzing a joint probability distribution of word annotations and image features [11], in which word probabilities are estimated using a multiple Bernoulli model and image feature probabilities are calculated using a non-parametric kernel density estimation. Based on the assumption that regions in an image can be represented with a small vocabulary of blobs, a cross-media relevance model is proposed in [12] to predict the probability of generating a word given the blobs in an image. Different from [12], regions in an image are described by a continuous-valued feature vector in [13]. A novel multi-label correlated Greens function approach is developed in