Abstract

With the rapid proliferation of Internet, it has become a great challenge to annotate a large number of objects manually, especially for the fashion domain where a massive collection of new products come up everyday. Moreover, the huge profits brought by the shopping on fashion products have motivated the multimedia search using the computer vision based techniques. Therefore, to save human labor, it is essential to develop an automatic tagging system (i.e., attribute prediction) for those fashion products in a variety of appearances. With the annotated attributes, it enables the online shoppers to perform the retrieval of their desired shoes. However, it fails to work on the shoe images from the daily life due to their large visual difference from the images from online stores. In this thesis, we study the shoe attribute prediction and retrieval system for the online product images as well as the daily life shoe photos. More specifically, we address the problem of in-store shoe retrieval as well the more challenging issue of the cross-scenario shoe retrieval guided by the semantic attributes of shoes. The works in this thesis can be summarized as below.

An in-store shoe retrieval system which allows the query images in the form of multi-view shoe images is firstly proposed, where each shoe is indexed by a list of part-aware shoe attributes. Given a set of multi-view shoe images, we first identify the viewpoint of each shoe image and a set of relevant view images are selected to estimate the value of each shoe attribute. To effectively predict the attributes which are part-aware, we incorporate the prior knowledge of the shoe structure under a certain viewpoint to learn a novel view-specific part localization

model, which localizes the shoe part from each of the relevant views. Experimental results demonstrate the effectiveness of the proposed system on a newly-built structured multi-view online shoe dataset.

Not limited to the attribute prediction and shoe retrieval for the online store shoe images, we also relax the constraint to perform the same task for the daily life photos with cluttered background, different scales, varied viewpoints, etc. A novel cross-domain shoe retrieval system is presented which aims to find the exactly same shoes given the query daily life shoe photos. More specifically, we propose the Semantic Hierarchy Of attributE Convolutional Neural Network (SHOE-CNN) with a newly designed loss function which systematically merges semantic attributes of closer visual appearances to avoid shoe images with the obvious visual differences being confused with each other. Moreover, a coarse-to-fine three-level feature representation is developed to effectively match the shoe images across different domains. The experimental results demonstrate the advantages of each component of our proposed system and a significant improvement over other baseline methods.

To further enable the retrieval of online store images with different viewpoints and address the failure cases with the viewpoint variation while at the same time improving the capability of differentiating the fine-grained details, we propose the feature embedding for shoes via a multi-task view- invariant convolutional neural network (MTV-CNN), the feature activations of which reflect the inherent similarity between any two shoe images. Specifically, we propose 1) the weighted triplet loss to reduce the feature distance between the same shoe in different scenarios; 2) a novel viewpoint invariant loss to reduce ambiguous feature representation from different views; 3) a novel definition of shoe style based on combinations of part-aware semantic shoe attributes and the corresponding style identification loss are presented; 4) the attribute-based hard negative and anchor images mining process to distinguish fine-grained differences. The experiments conducted on our newly collected dataset indicate that we are capable of not only returning the exactly same shoes or similar shoes but also different viewpoint ones.