Common Visual Pattern Discovery and Search

Zhenzhen Wang[1], Jingjing Meng[1], Tan Yu[2] and Junsong Yuan[1]
[1]School of Electrical and Electronic Engineering, [2]Interdisciplinary Graduate School
Nanyang Technological University, Singapore, 637553
E-mail: {zwang033, tyu008}@e.ntu.edu.sg, {jingjing.meng, jsyuan}@ntu.edu.sg

***Abstract*—Automatically discovering common visual patterns from images and videos is a useful but challenging task. On the one hand, the definition of visual patterns is rather ambiguous, it refers to the spatial composition of frequently occurring visual primitives which correspond to local features, semantic visual parts or visual objects. For example, the wheels and the body of a car could be seen as different visual primitives, while the whole car can also be seen as an individual visual primitive. On the other hand, there exhibit large variations in visual appearance and structures even within the same kind of visual pattern, which makes visual pattern discovery a very challenging task. However, since to distinguish different kinds of visual patterns from each other is a fundamental problem of many tasks in computer vision, such as pattern recognition/classification, object detection/localization, content-based image search, many studies have been introduce to solve the problem of visual pattern discovery in the literature. In this paper, we will revisit the representative studies on discovering visual patterns and discuss these methods from the view of local-feature-based and object-proposal-based visual patterns. The local-feature-based visual pattern discovery aims to mine the visual primitives that share similar spatial layout, while the semantic-patch-based visual pattern discovery aims to mine similar semantic patterns from the object proposals that are likely to contain an entire object. Then the extensive applications of visual pattern discovery are presented.**

I. Introduction

Visual pattern discovery aims to mine the re-occurring composition of visual primitives from a collection of images or videos even without manually labeled annotations [98], [81]. This topic recently draws increasing attention due to the fact that automatically summarizing the key content from a large body of visual data could be time- and labor-saving, especially in this big visual data era where there are millions of GB visual data being uploaded to Internet every day. This topic is also fundamental to many computer vision problems, such as image classification, content-based image retrieval, and object detection, since the common patterns could help to perceive and analyze the given image collections. Based on the commonalities of a specific pattern and the differences between different patterns, we can differentiate a dog from a cat, the foreground from the background, a red apple from a green one, and even the photo of a person at different age. Fig. 1 illustrates the general case of common pattern discovery.

However, to discover visual pattens from a random collection of images is quite a challenging task, in part because the definition of visual primitive is not as clear as in transaction and text data where usually the discrete elements are predefined. For example, the visual primitives could be semantic

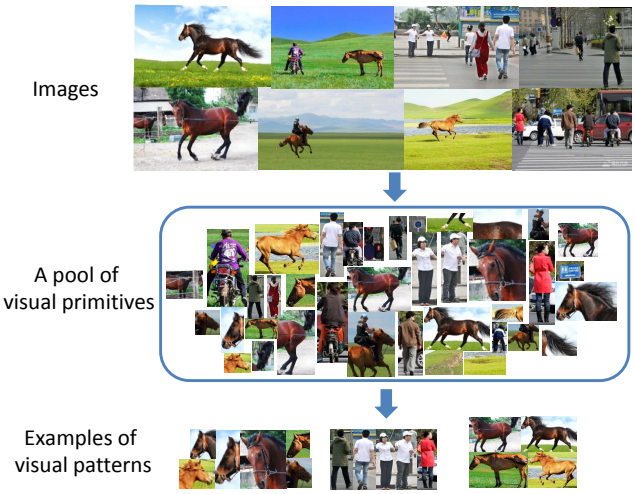

Fig. 1. The goal of the common pattern discovery is to mine the frequently occurring visual primitives from a collection of images.

visual parts as shown in Fig. 2: a bicycle is composed of two wheels (circles) and one triangle skeleton, each part of the bicycle could be seen as an individual visual primitive [27], [50]. With the development of techniques in extracting object proposals [69], [14], [12], it is also possible to crop an entire visual object, *i.e.*, the bicycle as a whole, from the image, thus we can also regard the whole bicycle as a visual primitive. Although the background may occur more frequently than the foreground object in some cases, *e.g.*, the sky and road in nature images, most of the researchers focus on the meaningful foreground objects. In this paper, we will revisit the representative studies on visual pattern discovery in terms of the local feature based methods and the object proposals based methods. Another challenge is that the visual primitives can be very diverse on their own. Large variations may be present in visual appearance and structure. In Fig. 2(b), we can see that the bicycle wheels could vary largely, not to mention the whole bicycle. Besides various illuminations and scales, the occlusion and distortion further present more difficulties in mining common visual patterns.

Extracting visual primitives from image collections and video data is the very first step for visual pattern mining, and good-quality visual primitives will definitely contribute to the mining results. Following our category of pattern discovery methods, *i.e.*, local feature based and semantic object proposal

Fig. 2. Examples of the visual primitives.

based methods, we briefly review some representative studies on collecting local primitive regions, and on extracting object proposals. For the former, many local feature detectors [82] are popularly used to obtain visual primitives, such as blobs. Normalized Cuts [72], which is firstly proposed to solve the perceptual grouping problem for image segmentation, can also be used to collect primitive regions. The deformable primitive models [26], which is extensively used in object detection tasks, can be adopted to generate object primitives. For the latter, there are many methods on obtaining object proposals, such as Selective Search [83], Randomized Prim's [57], EdgeBoxes [115], and Bing [15]. These methods usually generate the candidate proposals with scores indicating the probability of containing an object, thus can significantly reduce the number of candidate segmentations compared to the dense framework, *e.g.* sliding window.

The main difference of the two categories is that for the former, the visual primitives which represent local interest points or regions are collected by randomly decomposing the images, then some post-processing steps will be conducted to select the common spatial structure primitives and the common patterns composed of these primitives; while the object proposals are usually generated from pre-trained models and are more likely to correspond to the whole objects, then the post-processing steps are to group these object proposals and discover the frequently occurred patterns. Intuitively, the pattern discovery methods based on the local primitives would involve heavier computational cost than the methods based on the object proposals. However, since the object proposals tend to contain a whole object which can exhibit large visual appearance variations than the local visual primitives, it would be more difficult to mine common patterns from the object proposals.

II. Local Visual Primitives based Pattern Discovery

Given a set of images and each of which is characterized by a number of local visual primitives, lots of studies have been published on mining the visual patterns from them in the past decade. Most previous methods based on the local visual primitives can be roughly divided into bottom-up and top-down approaches. In the former, different models are designed to gradually group the visual primitives extracted from image collections, then the frequently occurred spatial composition of visual primitives are selected as visual patterns. In the latter, the models are directly built on the labeled images and the segmentations to learn the parameters, so that the trained model could be used to infer the common visual patterns from unseen images.

A. Bottom-up Approach

One of the most popularly used bottom-up approaches is to formulate the common pattern discovery as the problem of sub-graph mining [78], [38], [79], [30]. In these methods, each image can be represented as a graph where the visual primitives correspond to vertices and the similarities between visual primitives correspond to edges if any. Some variations can be found in literatures. For example, Liu and Yan [53] extend the above mentioned sub-graph mining on an individual image to mine the common patterns from a pair of images. In [110], [113], a novel cohesive subgraph mining method is proposed to discover thematic patterns from a single video. Unlike pattern mining from images, the resulting visual primitives should be spatio-temporally collocated. To this end, an algorithm is proposed to find the topical objects by maximizing the overall mutual information scores. An image can be represented as a tree characterized by image segmentation in different scales. The larger segmentation can be further decomposed into small ones as its child nodes, then the maximally matching subtrees correspond to the common patterns among a given image collection [79].

Frequent item set mining algorithms (FIM) [73] are also widely applied to the bottom-up methods. FIM is originally designed for searching frequent sets from supermarket transaction data, it can be easily tailored to frequent pattern mining by treating visual primitives as transaction items, and an image as collection of items from a consumer [93], [99], [103], [104]. In past years, many researchers put efforts on extending the traditional FIM methods to visual pattern mining. For example, in order to capture invariant relative positions of a pair of objects, Hsu *et al.* [41] adopt the Apriori algorithm for mining patterns composed of objects with stable relative position. In [76], Sivic and Zisserman propose a clustering based algorithm to group the visual primitives that exhibit typical