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

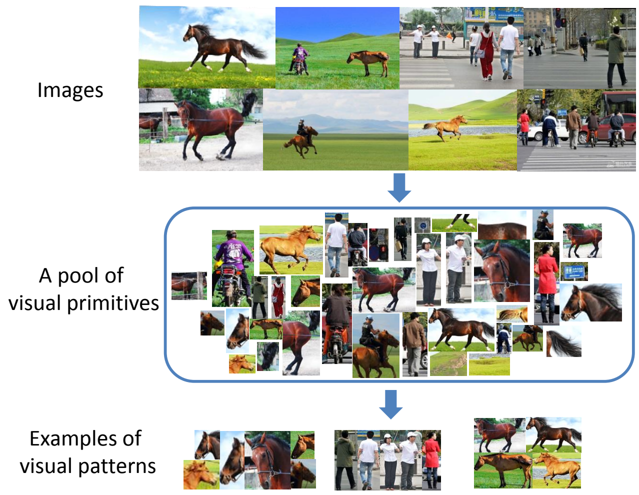

Fig. 1. The goal of the common pattern discovery is to mine the frequently occurring visual primitives from a collection of images.

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 prototypes; Yuan *et al.* [101], [103], [102] adopt the FP-growth based method [65] to identify the frequently occurring visual patterns.

Before mining the visual patterns using FIM algorithms, we need to build transaction data which are usually represented as binary vectors with 1 indicating the present of a specific item. To obtain the transaction database, similar visual primitives are grouped as visual words, then a visual vocabulary is composed of these visual words. To narrow the errors between visual words and visual primitives, an unsupervised context-aware clustering method is proposed in [100] to improve the quality of visual words, so that the results of the discovered visual patterns can be further improved. Later, Wang *et al.* [86] are inspired by the context-aware clustering algorithm and propose the multiple-feature based clustering. Although the quantization error of transaction data obtained from visual primitives can be significantly reduced by the above mentioned methods, to sidestep the building of transaction database can also contribute to a better performance. For example, in [109] and [104], a multilayer candidate pruning based algorithm is introduced and in reference [99], a spatial random partition algorithm is proposed.

B. Top-down Approach

The top-down methods directly build models based on tools such as pLSA [37] and LDA [9] from text analysis literatures, for frequently occurring visual patterns in a given image collection, so that the common visual patterns can be discovered from unseen image collections which follow the same distribution as training images. Liu and Chen [52] introduce a combined framework which exploits the pLSA and a motion model based on PDA filter [6] to mine the common objects from videos. The LDA model is also used to discover visual object classes from image collections [71]. Later, Sivic *et al.* [75] introduce the hierarchical LDA (hLDA) model to mine the hierarchical structure of visual patterns. Both [71] and [75] require to randomly segment the images several times to generate a pool of segmentations, then the object topics or the object hierarchies are to discovered from the pool. Unlike [71], [75], which are working based on segmentations, reference [3] introduces a combined model which leverages a hybrid parametric-nonparametric model and the LDA for discovering objects and segmentations with specified category.

Since the spatial relationships between visual primitives are also key to visual patterns discovery, a spatial LDA (sLDA) algorithm is proposed in [89] to solve vision problems. Different from traditional LDA model used in language problems, which only considers the presence or absence of visual words, the sLDA could incorporate the spatial relationships between visual words. Similarly, a geometric LDA (gLDA) is introduced later in [67] to better encode the homographic geometric relationships among visual words. Besides finding the visual patterns, the LDA based models can also be extended to pixel-level image segmentation and even recognize different object and scene categories [11]. There are also studies on exploring better topic models using different priors [8], [4], similar to that the word "disease" will appear more often than "car" in a document about health, the prior knowledge about image collections can also be used in patterns discovery. Markov chain is applied to model the topics of words for language problems [32]. Then Zhao *et al.* [111] leverage the Gaussian Markov chain to model word co-occurrence prior for common object discovery from videos. It is then incorporated in traditional LDA model so that the bottom-up priors and the top-down probabilistic model could benefit each other.

Instead of incorporating spatial relationships between visual words as constraints like sLSA and gLDA models, one can also directly model the spatial structure using graphs or hierarchical tree. For example, in [38] an algorithm for automatically discovering spatial patterns is introduced. Specifically, the proposed algorithm models the spatial pattern as a mixture of probabilistic parametric attributed relational graphs with each node of the graph corresponding to the segmentation from images, then the parameters of the model are optimized using expectation-maximization (EM) algorithm. The main difference between graph and tree is that the latter can build a hierarchical and more complex topological structure. Therefore, Todorovic and Ahuja [79] characterize each image as a tree with multiple-scale image segmentations as its nodes, and the common patterns are discovered by maximally matching subtrees from the image collection.

III. Object Proposals based Pattern Discovery

Although any frequently occurring structure of visual primitives could construct the patterns, *e.g.*, the sky and cloud, the majority of the researchers focus on the re-occurring objects across images, *i.e.*, the common objects. Different from the methods based on local visual primitives which iteratively merge the image patches until the larger objects are found, recent studies tend to directly generate the patches from a collection of images that contain the whole objects, *i.e.*, the object proposals [40], [39]. To differentiate the objects from non-objects can also be seen as common pattern discovery since there must be some common patterns shared by objects, then to group the object proposals containing the same object class is to mine more specific patterns. Thus in this section we will firstly revisit some representative studies on generating object proposals, then on mining common objects.

A. Generating object proposals

Essentially, common object discovery is to simultaneously distinguish the objects from backgrounds and group similar objects together. It is easy to understand that different categories of objects are diverse in their feature space, but how can we distinguish the objects and backgrounds? Based on the observation that objects exhibit common visual characteristics, researchers can design methods based on their observations of visual properties [84], [83], [69], [12], [13], [23], [24], [5], [115], or one can directly train a method based on the given labeled samples [57], [14], [1], [2], [15], [108]. These methods are usually classified into methods based on grouping and window scoring according to whether outputting scores for candidate windows or not [39]. Specially, proposal generating methods based on grouping try to extract the segments that tend to contain the whole objects *e.g.*, Selective Search [83], Randomized Prim's [57], Chang [14], while window scoring proposal methods will generate object proposals with scores indicating the probability of containing an object, *e.g.*, Objectness [1], [2], Bing [15], EdgeBoxes [115]. There are also other proposal methods not falling into the two categories. For example, Multibox [77], [25] takes advantage of the powerful learning ability of CNN, and train a CNN that can directly output the bounding boxes of proposals.

B. Traditional object discovery methods

After obtaining the object proposals, many classic methods designed for pattern discovery from visual primitives, such as sub-graph mining [63], [48], [106], K-medoids clustering [46], [95], and dense correspondence [36], could be applied to common objects discovery. To mine common objects by graph-based algorithms, Kapil and Maheshkumar [34] propose an unsupervised graph based framework where each of the object proposals corresponds to a vertex of the graph and the similarities between object proposals correspond to the edges of the graph. The object detection result is obtained by maximizing weighted path in the graph. Similarly, Federico *et al.* [66] exploit a fully connected spatial-temporal graph built over object proposals, where they formulate the video segmentation problem as a minimization of an energy function defined over the graph, so that the long range relations in videos can be modeled and the information between both spatially and temporally distant object proposals can be exchanged.

K-medoids clustering based methods are also widely used in common object discovery, in which K object proposals are selected as clustering centroids, and the rest of the object proposals are associated to the closet centroids. Yu *et al.* [95] apply the K-mediods clustering to the object proposals extracted from reference images to accelerate content-based image search. To reduce the cost of the configuration, Elhamifar *et al.* [22] improve K-medoids clustering by representing each sample using multiple centroids, then the effectiveness of the improved algorithm is demonstrated in the problem of image classification and video summarization. However, K-mediods clustering may not be appropriate in unsupervised object discovery where the number of common object classes is unknown. Therefore, dense correspondences between object proposals can be used in such a case, and the images which do not contain any common objects can also be identified. In addition, higher saliency can act as an indicator of present of common object patterns, Rubinstein *et al.* [70] use dense correspondences between images to discover and segment out common objects from large and diverse image collections. In [35], a semantic flow approach, termed proposal flow, is proposed to establish reliable correspondences using object proposals. The proposal flow could generate a reliable semantic flow between a pair of similar images using local and geometric consistency constraints among object proposals. Inspired by proposal flow, Minsu *et al.* [16] present an unsupervised framework based on part-based matching with object proposals for common object discovery and localization from a noisy image collection.

C. End-to-End Neural Network

Recent years, CNN has been extensively used in computer vision and achieved state-of-the-art performance in most of the problems, *e.g.*, object detection, tracking, and classification, and can even execute some novel tasks such as style transfer [55] that traditional methods cannot do. One of the drawbacks of CNN is that it needs a large body of labeled training samples to learn numerous parameters, thus it is difficult to be applied to the common pattern discovery problem which usually do not provide or only provide a few labeled samples. As an unsupervised network, autoencoder has been used to remove outlier images from an image collection [92]. Inspired by the observation that the reconstruction error of inliers is smaller than outliers, Xia *et al.* [92] gradually train the network so that the images containing the common object could be identified. Given a set of images containing the same object class, [7] and [44] train CNN with the image-level label so that the finely tuned CNN could score the object proposals containing the common objects higher. Li *et al.* [51] propose a novel CNN architecture to discover common visual patterns. The insight of their study is that the trained convolution layers are able to capture local texture patterns of a given image set, thus the activations of filters in a CNN could be leveraged to automatically discover patterns.

IV. Applications

Visual patterns are basic visual elements that can capture the spatial layout of re-occurring visual primitives. The semantic visual patterns can be showed in different level, such as lines, dots, wheels, horses, etc. The study of visual pattern discovery has important implications on a series of problems, such as image and instance search, recognition, and video analysis.

Visual Search. Visual search from image collections can be roughly categorized into content-based/instance search and image search. The instance search cares more about whether the reference images contain similar object to the query while image search prefers to retrieve references globally similar to the query. Yu *et al.* [95] propose a Fuzzy Objects Matching (FOM) framework for instance search from image collections, whose main contribution is applying the k-mediods to the object proposals extracted from all reference images so that the search complexity could be reduced. Later, they extend the instance search to videos and propose an hierarchical object prototype encoding (HOPE) model to accelerate the object instance search in videos [96]. Zhang *et al.* [107] introduce a unified framework for image search, which combines a geometric visual vocabulary for better encoding the spatial relationships of visual primitives and a learned semantic-aware distance metric for obtaining the semantic context. Jiang *et al.* [43] incorporate the spatial context of image patches from references and construct visual phrase for better matching. Compared with matching visual words, *i.e.*, the individual patches separately, the utilization of spatial context could be more robust.

Object Categorization. The fundamental problem of image and object categorization is to find images or objects that share similar visual patterns. Previous methods in the literature can be roughly categorized as "bag-of-words" models and part-based models. To improve the performance of "bag-of-words" based models, efforts can be put into improving the quality of visual dictionary [101], [90], [61], [58], so that a more separable feature representation can be obtained; or a more discriminative classifier, such as SVM, which can be applied to better distinguish different categories [85], [18]; or a more powerful generative model such as pLSA and LDA that can be built to discover object categories [10], [54], [68]. For part-based matching models, to densely match all the segmentations from a pair of images would be very time consuming, thus many studies explore efficient models to mine common objects, such as building graph based models on image patches [66], [106] or tree structure [79], [38]. A better distance metric is also key to improve the matching results [31], [49].

Video Analysis. Visual pattern mining also has numerous applications in video analysis. Since each video is simply a set of image frames, discovering visual patterns from the video frames can help to find the repetitive image regions appearing frequently throughout the video. This information can be used to perform efficient video summarization, compression, indexing and search [59], [97]. Moreover, these frequently appearing visual patterns provide important clues to discovering the objects that consistently appear throughout a video sequence [104], [112], [111], [56], [114]. Besides within the frames of a single video, discovering visual patterns within a collection of videos can benefit the video object co-localization tasks, which is to find the objects appearing frequently in a video dataset [45], [47], [42], [80], [105], [88], [87], [29], [28], [21]. Furthermore, when the pattern mining is applied to motion domain to perform motion pattern mining, it can be used to discover the repetitive actions within a video collection [17], [33], [64], [20], [19], [60], [62], [74], [91], [94].

V. Conclusions

Visual pattern discovery is a fundamental problem in computer vision. The essence of many visual studies, such as object recognition and visual search, is to mine the commonalities and differences among image patches of interest, so that the representative and discriminative patterns can be obtained, then one can group or differentiate any two image regions according to the understanding of the given image collections. One of the difficulties of pattern discovery lies in the ambiguous concept of pattern, which can vary at different semantic levels, for example, lines, wheels, cars, etc. Therefore, in this survey, we revisit and categorize previous methods from two views, *i.e.*, visual primitives and semantic object proposals. We also briefly discuss the applications of visual pattern discovery.

Although tremendous progress has been made in visual pattern discovery during past decades, much work still needs to be done in this area to help us better perceive our visual world since understanding and modeling visual patterns are the key of perceiving our physical world.

Acknowledgment

The authors thank Jiong Yang and Weixiang Hong for their valuable discussion and support.

References

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 73–80. IEEE, 2010.

[2] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2189–2202, 2012.

[3] Marco Andreetto, Lihi Zelnik-Manor, and Pietro Perona. Unsupervised learning of categorical segments in image collections. *IEEE transactions on pattern analysis and machine intelligence*, 34(9):1842–1855, 2012.

[4] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32. ACM, 2009.

[5] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014.

[6] Yaakov Bar-Shalom, Fred Daum, and Jim Huang. The probabilistic data association filter. *IEEE Control Systems*, 29(6), 2009.

[7] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[8] David M Blei and John D Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[10] Anna Bosch, Xavier Muñoz, and Robert Martí. Which is the best way to organize/classify images by content? *Image and vision computing*, 25(6):778–791, 2007.

[11] Liangliang Cao and Li Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[12] Joao Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3241–3248. IEEE, 2010.

[13] Joao Carreira and Cristian Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1312–1328, 2012.

[14] Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. Fusing generic objectness and visual saliency for salient object detection. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 914–921. IEEE, 2011.

[15] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3286–3293, 2014.

[16] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2015.

[17] Wen-Sheng Chu, Feng Zhou, and Fernando De la Torre. Unsupervised temporal commonality discovery. In *Proceedings of the European Conference on Computer Vision*, pages 373–387. Springer, 2012.

[18] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[19] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Articulated motion discovery using pairs of trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2151–2160. IEEE, 2015.

[20] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Behavior discovery and alignment of articulated object classes from unstructured video. *International Journal on Computer Vision*, pages 1–23, 2016.

[21] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Discovering the physical parts of an articulated oclass from multiple videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 714–723, 2016.

[22] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In *Advances in Neural Information Processing Systems*, pages 19–27, 2012.

[23] Ian Endres and Derek Hoiem. Category independent object proposals. *Computer Vision–ECCV 2010*, pages 575–588, 2010.

[24] Ian Endres and Derek Hoiem. Category-independent object proposals with diverse ranking. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):222–234, 2014.

[25] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2154, 2014.

[26] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[27] Sanja Fidler and Ales Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[28] Huazhu Fu, Dong Xu, Bao Zhang, and Stephen Lin. Object-based multiple foreground video co-segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173. IEEE, 2014.

[29] Huazhu Fu, Dong Xu, Bao Zhang, Stephen Lin, and Rabab Kreidieh Ward. Object-based multiple foreground video co-segmentation via multi-state selection graph. *T-IP*, 24(11):3415–3424, 2015.

[30] Jizhou Gao, Yin Hu, Jinze Liu, and Ruigang Yang. Unsupervised learning of high-order structural semantics from images. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2122–2129. IEEE, 2009.

[31] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1458–1465. IEEE, 2005.

[32] Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. Hidden topic markov models. In *Artificial intelligence and statistics*, pages 163–170, 2007.

[33] Jiaming Guo, Zhuwen Li, Loong-Fah Cheong, and Steven Zhiying Zhou. Video co-segmentation for meaningful action extraction. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2232–2239, 2013.

[34] Kapil Gupta and Maheshkumar Kolekar. Unsupervised graph based video object extraction. 2015.

[35] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3475–3484, 2016.

[36] Tal Hassner and Ce Liu. *Dense Image Correspondences for Computer Vision*. Springer, 2016.

[37] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1):177–196, 2001.

[38] Pengyu Hong and Thomas S Huang. Spatial pattern discovery by learning a probabilistic parametric model from multiple attributed relational graphs. *Discrete Applied Mathematics*, 139(1):113–135, 2004.

[39] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? *IEEE transactions on pattern analysis and machine intelligence*, 38(4):814–830, 2016.

[40] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. How good are detection proposals, really? *arXiv preprint arXiv:1406.6962*, 2014.

[41] Wynne Hsu, Jing Dai, and Mong Li Lee. Mining viewpoint patterns in image databases. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 553–558. ACM, 2003.

[42] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Cats: Co-saliency activated tracklet selection for video co-localization. In *Proceedings of the European Conference on Computer Vision*, pages 187–202. Springer, 2016.

[43] Yuning Jiang, Jingjing Meng, and Junsong Yuan. Randomized visual phrases for object search. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3100–3107. IEEE, 2012.

[44] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. *arXiv preprint arXiv:1704.05188*, 2017.

[45] Armand Joulin, Kevin D Tang, and Fei-Fei Li. Efficient image and video co-localization with frank-wolfe algorithm. In *Proceedings of the European Conference on Computer Vision*, pages 253–268. Springer, 2014.

[46] Leonard Kaufman and Peter Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.

[47] Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Unsupervised object discovery and tracking in video collections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3173–3181. IEEE, 2015.

[48] Eugene L Lawler. *Combinatorial optimization: networks and matroids*. Courier Corporation, 2001.

[49] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[50] Congcong Li, Devi Parikh, and Tsuhan Chen. Automatic discovery of groups of objects for scene understanding. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2735–2742. IEEE, 2012.

[51] Hongzhi Li, Joseph G Ellis, Lei Zhang, and Shih-Fu Chang. Patternnet: Visual pattern mining with deep neural network. *arXiv preprint arXiv:1703.06339*, 2017.

[52] David Liu and Tsuhan Chen. A topic-motion model for unsupervised video object discovery. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[53] Hairong Liu and Shuicheng Yan. Common visual pattern discovery via spatially coherent correspondences. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1609–1616. IEEE, 2010.

[54] Zhiwu Lu and Horace HS Ip. Image categorization with spatial mismatch kernels. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 397–404. IEEE, 2009.

[55] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. *arXiv preprint arXiv:1703.07511*, 2017.

[56] Ye Luo, Gangqiang Zhao, and Junsong Yuan. Thematic saliency detection using spatial-temporal context. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 347–353, 2013.

[57] Santiago Manen, Matthieu Guillaumin, and Luc Van Gool. Prime object proposals with randomized prim's algorithm. In *Proceedings of the IEEE international conference on computer vision*, pages 2536–2543, 2013.

[58] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

[59] Jingjing Meng, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. From keyframes to key objects: Video summarization by representative object proposal selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1039–1048, 2016.

[60] Pascal Mettes, Jan C van Gemert, and Cees GM Snoek. Spot on: Action localization from pointly-supervised proposals. In *Proceedings of the European Conference on Computer Vision*, pages 437–453. Springer, 2016.

[61] Krystian Mikolajczyk and Cordelia Schmid. An affine invariant interest point detector. *Computer VisionECCV 2002*, pages 128–142, 2002.

[62] Ehsan Adeli Mosabbeb, Ricardo Cabral, Fernando De la Torre, and Mahmood Fathy. Multi-label discriminative weakly-supervised human activity recognition and localization. In *Proceedings of Asian Conference on Computer Vision*, pages 241–258. Springer, 2014.

[63] Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

[64] Konstantinos Papoutsakis, Costas Panagiotakis, and Antonis A Argyros. Temporal action co-segmentation in 3d motion capture data and videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[65] Jian Pei, Jiawei Han, and Laks VS Lakshmanan. Mining frequent itemsets with convertible constraints. In *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 433–442. IEEE, 2001.

[66] Federico Perazzi, Oliver Wang, Markus Gross, and Alexander Sorkine-Hornung. Fully connected object proposals for video segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3227–3234, 2015.

[67] James Philbin, Josef Sivic, and Andrew Zisserman. Geometric latent dirichlet allocation on a matching graph for large-scale image datasets. *International journal of computer vision*, 95(2):138–153, 2011.

[68] Pedro Quelhas, Florent Monay, J-M Odobez, Daniel Gatica-Perez, Tinne Tuytelaars, and Luc Van Gool. Modeling scenes with local descriptors and latent aspects. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 883–890. IEEE, 2005.

[69] Pekka Rantalankila, Juho Kannala, and Esa Rahtu. Generating object segmentation proposals using global and local search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2417–2424, 2014.

[70] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1939–1946, 2013.

[71] Bryan C Russell, William T Freeman, Alexei A Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1605–1614. IEEE, 2006.

[72] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[73] Julianna Katalin Sipos. Frequent item set mining methods.

[74] Parthipan Siva and Tao Xiang. Weakly supervised action detection. In *Proceedings of British Machine Vision Conference*, volume 2, page 6, 2011.

[75] Josef Sivic, Bryan C Russell, Andrew Zisserman, William T Freeman, and Alexei A Efros. Unsupervised discovery of visual object class hierarchies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[76] Josef Sivic and Andrew Zisserman. Video data mining using configurations of viewpoint invariant regions. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2004.

[77] Christian Szegedy, Scott Reed, Dumitru Erhan, Dragomir Anguelov, and Sergey Ioffe. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*, 2014.

[78] Hung-Khoon Tan and Chong-Wah Ngo. Localized matching using earth movers distance towards discovery of common patterns from small image samples. *Image and Vision Computing*, 27(10):1470–1483, 2009.

[79] Sinisa Todorovic and Narendra Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2158–2174, 2008.

[80] Yi-Hsuan Tsai, Guangyu Zhong, and Ming-Hsuan Yang. Semantic co-segmentation in videos. In *Proceedings of the European Conference on Computer Vision*, pages 760–775. Springer, 2016.

[81] Tinne Tuytelaars, Christoph H Lampert, Matthew B Blaschko, and Wray Buntine. Unsupervised object discovery: A comparison. *International journal of computer vision*, 88(2):284–302, 2010.

[82] Tinne Tuytelaars, Krystian Mikolajczyk, et al. Local invariant feature detectors: a survey. *Foundations and trends® in computer graphics and vision*, 3(3):177–280, 2008.

[83] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[84] Koen EA Van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders. Segmentation as selective search for object recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1879–1886. IEEE, 2011.

[85] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[86] Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. Combining feature context and spatial context for image pattern discovery. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 764–773. IEEE, 2011.

[87] Le Wang, Gang Hua, Rahul Sukthankar, Jianru Xue, Zhenxing Niu, and Nanning Zheng. Video object discovery and co-segmentation with eweak supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[88] Le Wang, Gang Hua, Rahul Sukthankar, Jianru Xue, and Nanning Zheng. Video object discovery and co-segmentation with extremely weak supervision. In *Proceedings of the European Conference on Computer Vision*, pages 640–655. Springer, 2014.

[89] Xiaogang Wang and Eric Grimson. Spatial latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1577–1584, 2008.

[90] Markus Weber, Max Welling, and Pietro Perona. Unsupervised learning of models for recognition. *Computer Vision-ECCV 2000*, pages 18–32, 2000.

[91] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Towards weakly-supervised action localization. *arXiv preprint arXiv:1605.05197*, 2016.

[92] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1519, 2015.

[93] Yan Xie and S Yu Philip. Max-clique: A top-down graph-based approach to frequent pattern mining. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 1139–1144. IEEE, 2010.

[94] Donghun Yeo, Bohyung Han, and Joon Hee Han. Unsupervised co-activity detection from multiple videos using absorbing markov chain. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3662–3668, 2016.

[95] Tan Yu, Yuwei Wu, Sreyasee Das Bhattacharjee, and Junsong Yuan. Efficient object instance search using fuzzy objects matching. In *AAAI*, pages 4320–4326, 2017.

[96] Tan Yu, Yuwei Wu, and Junsong Yuan. Hope: Hierarchical object prototype encoding for efficient object instance search in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[97] Tan Yu and Junsong Yuan. Compressive quantization for fast object instance search in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[98] Junsong Yuan. Discovering visual patterns in image and video data: Concepts, algorithms, experiments. Saarbrcken, Germany: VDM Verlag Dr. Mller, 2011.

[99] Junsong Yuan and Ying Wu. Spatial random partition for common visual pattern discovery. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[100] Junsong Yuan and Ying Wu. Context-aware clustering. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[101] Junsong Yuan and Ying Wu. Mining visual collocation patterns via self-supervised subspace learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):334–346, 2012.

[102] Junsong Yuan, Ying Wu, and Ming Yang. Discovery of collocation patterns: from visual words to visual phrases. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[103] Junsong Yuan, Ying Wu, and Ming Yang. From frequent itemsets to semantically meaningful visual patterns. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 864–873. ACM, 2007.

[104] Junsong Yuan, Gangqiang Zhao, Yun Fu, Zhu Li, Aggelos K Katsaggelos, and Ying Wu. Discovering thematic objects in image collections and videos. *IEEE Transactions on Image Processing*, 21(4):2207–2219, 2012.

[105] Dong Zhang, Omar Javed, and Mubarak Shah. Video object co-segmentation by regulated mweight cliques. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 551–566. Springer, 2014.

[106] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Mining and-or graphs for graph matching and object discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 55–63, 2015.

[107] Shiliang Zhang, Qi Tian, Gang Hua, Wengang Zhou, Qingming Huang, Houqiang Li, and Wen Gao. Modeling spatial and semantic cues for large-scale near-duplicated image retrieval. *Computer Vision and Image Understanding*, 115(3):403–414, 2011.

[108] Ziming Zhang, Jonathan Warrell, and Philip HS Torr. Proposal generation for object detection using cascaded ranking svms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1497–1504. IEEE, 2011.

[109] Gangqiang Zhao and Junsong Yuan. Mining and cropping common objects from images. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 975–978. ACM, 2010.

[110] Gangqiang Zhao and Junsong Yuan. Discovering thematic patterns in videos via cohesive sub-graph mining. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 1260–1265. IEEE, 2011.

[111] Gangqiang Zhao, Junsong Yuan, and Gang Hua. Topical video object discovery from key frames by modeling word co-occurrence prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1602–1609, 2013.

[112] Gangqiang Zhao, Junsong Yuan, and Yang Jiong Hua, Gang. Topical video object discovery from key frames by modeling word co-occurrence prior. *IEEE Transactions on Image Processing*, 2015.

[113] Gangqiang Zhao, Junsong Yuan, Jiang Xu, and Ying Wu. Discovering the thematic object in commercial videos. *IEEE MultiMedia*, 18(3):56–65, 2011.

[114] Gangqiang Zhao, Junsong Yuan, Jiang Xu, and Ying Wu. Discovering the thematic object in commercial videos. *IEEE MultiMedia*, 18(3):56–65, 2011.

[115] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.