Advanced Review

Visual pattern discovery in image and video data: a brief survey

Hongxing Wang, Gangqiang Zhao and Junsong Yuan*

In image and video data, visual pattern refers to re-occurring composition of visual primitives. Such visual patterns extract the essence of the image and video data that convey rich information. However, unlike frequent patterns in transaction data, there are considerable visual content variations and complex spatial structures among visual primitives, which make effective exploration of visual patterns a challenging task. Many methods have been proposed to address the problem of visual pattern discovery during the past decade. In this article, we provide a review of the major progress in visual pattern discovery. We categorize the existing methods into two groups: bottom-up pattern discovery and top-down pattern modeling. The bottom-up pattern discovery method starts with unordered visual primitives followed by merging the primitives until larger visual patterns are found. In contrast, the top-down method starts with the modeling of visual primitive compositions and then infers the pattern discovery result. A summary of related applications is also presented. At the end we identify the open issues for future research.

How to cite this article:
WIREs Data Mining Knowl Discov 2014, 4:24–37. doi: 10.1002/widm.1110

INTRODUCTION

Similar to frequent patterns in transaction data, visual patterns are compositions of visual primitives that appear frequently in image and video data.[1,2] The visual primitives that construct visual patterns can be very diverse, e.g., local image patches, semantic visual parts, or visual objects. As shown in Figure 1, the visual pattern in image or video data can be a texton that captures the repetitiveness of image texture,[6] e.g., the 'double-G' pattern in a Gucci bag; an abstract object model that describes its composition of visual parts,[7] e.g., a face pattern composed of two eyes, a nose, and a mouth; a scene layout pattern that captures the key objects which compose the scene,[8] e.g., a bedroom including a bed, a lamp, and so on; or a human action that describes postures and motions of human body, e.g., a bent-leg layover spin action showing by upturning the torso and bending the free leg. Such visual patterns are ubiquitous in images and videos. Just like the perception of repeated structures is well-nigh fundamental to the understanding the world around us,[9] the recognition of visual patterns is essential to the understanding of image and video data. In practice, visual patterns can be used to model images and videos, which have extensive applications in image and video analysis, such as image search, object categorization, video summarization, and human action recognition. It therefore offers an interesting, practical, but challenging task to mine visual patterns from images and videos.

Although frequent pattern mining has been well studied in data mining community,[10] the existing frequent pattern mining methods cannot be applied to image and video data directly. This is because the visual content variations and complex spatial structures among visual data make the problem of visual pattern discovery more challenging. Therefore, before mining visual patterns, it is required to extract stable visual primitives from image or video data. To obtain visual primitives, many local feature detectors have been proposed.[11] Segmentation methods, e.g., normalized cuts,[12] can be used to collect primitive regions. Object detection methods, e.g., deformable part models,[13] provide object primitives appearing in image or video data. Once we have visual primitives, we can encode their appearance using

*Correspondence to: jsyuan@ntu.edu.sg

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Conflict of interest: The authors have declared no conflicts of interest for this article.

FIGURE 1 | Diverse visual patterns: (a) the repetitive 'double-G' textures generate the texton patterns in a Gucci bag; (b) two eyes, a nose, and a mouth sketch a face pattern. Source: Images are from Caltech 101 dataset (Ref 3); (c) a bed, a lamp and so on usually make up a bedroom. Source: Images are from MIT Indoor dataset (Ref 4); (d) upturning of the torso and bending of the free leg together show the bent-leg layover spin action (Ref 5).

FIGURE 2 | Preprocessing of image and video data.

feature descriptors.[14] Instead of describing visual primitives using raw features, we can also use clustering method, e.g., *k*-means, to further quantize feature descriptors into discrete visual words. After that, each visual primitive can be identified by the corresponding visual word. Then an image can be described by a 'bag-of-visual-words'. We summarize the preprocessing of image or video data in Figure 2.

In the past decade, there have been increasing efforts to address visual pattern discovery in the literature. The aim of this article is to review recent work and provide an overview of this topic. We categorize the visual pattern discovery methods into two groups: bottom-up and top-down methods. The bottom-up pattern discovery methods start with visual primitives and then merge these primitives until the larger visual patterns are found. The basic idea is shown in Figure 3(a). First, each image is decomposed into a number of visual primitives. Then, the visual primitives are quantized into visual words (colored in blue) by clustering. After that, by investigating frequent visual word configurations in image spatial space, two types of word co-occurrence compositions, i.e., visual patterns {'cross', 'star'} and {'parallelogram', 'diamond', 'trapezoid'} are found. Finally, we locate all instances of both types of visual patterns. In contrast, the top-down methods start with the modeling of images and visual patterns and then infer the pattern discovery result. Figure 3(b) illustrates the top-down method by using the latent Dirichlet allocation (LDA) to model images and visual patterns.[15] The basic idea is that images are represented as mixtures over visual patterns, where each pattern is characterized by a distribution over visual words. This is similar to describing a document by mixtures of topics, where each topic has its own word distribution. The pattern discovery is achieved by inferring the posterior distribution of visual pattern mixture variable given an image. In this survey, we summarize the representative work of visual pattern discovery in Table 1. The datasets used in the corresponding work are also listed. Meanwhile, we organize our discussion into three parts: bottom-up pattern mining methods, top-down pattern mining methods, and applications of visual pattern discovery. In section *Conclusion and Outlook*, we conclude this study.

BOTTOM-UP PATTERN MINING

Classic frequent itemset mining (FIM) methods[10] provide off-the-shelf bottom-up techniques for pattern discovery from transaction data and inspire early research on visual pattern discovery. However, the performance of FIM-based methods heavily depends on the quality of transaction data. Thus more general strategies have been proposed to avoid the generation of transactions for image/video data mining, e.g., frequent pattern counting by visual primitive matching. Owing to modeling sophisticated spatial structures among visual primitives, many graph-based pattern mining methods have also been proposed.

Classic FIM Methods for Visual Pattern Discovery

Apriori,[85] frequent pattern growth (FP-growth)[86] and clustering are among the classic methods in FIM.[10] To leverage FIM algorithms for visual pattern discovery, one can build transaction data in local spatial neighborhoods of visual primitives. To be

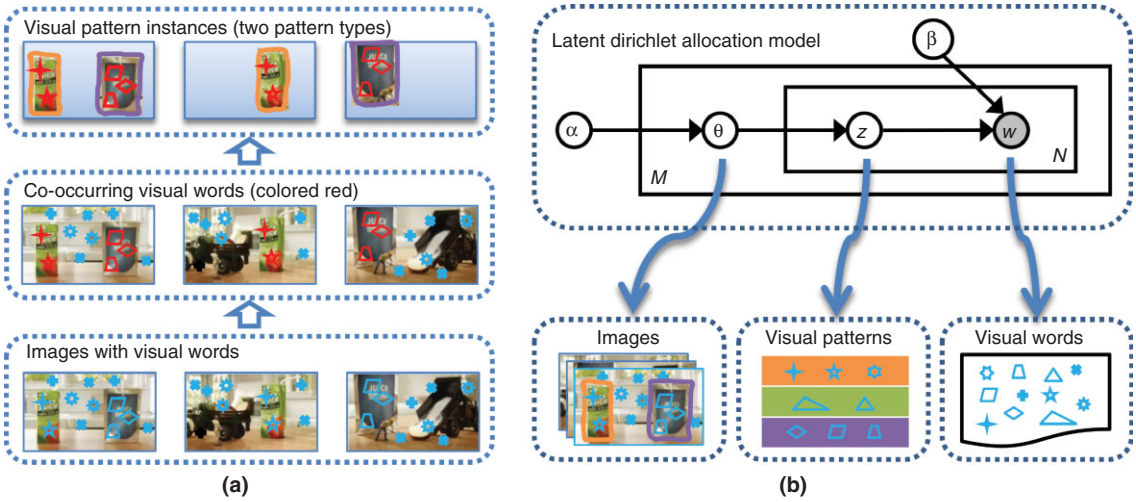

FIGURE 3 | Bottom-up (a) and top-down (b) visual pattern discovery.

specific, a transaction can be built to represent a spatial neighborhood of a visual primitive with a binary vector that indicates whether a visual word is present or not within this neighborhood. As an image can generate a number of transactions, the classic FIM methods can be applied to visual pattern discovery.

Hsu et al.[17] have early adopted the Apriori algorithm in order to discover viewpoint patterns that capture invariant relationships among objects. Quack et al.[27] mine frequent spatial configurations of visual primitive patches using the Apriori algorithm.[85] Lee et al.[29] also utilize the Apriori algorithm to discover spatial association patterns from image data. To identify closed frequent visual patterns, Yuan et al.[63,87,88] apply the FP-growth algorithm.[86] Sivic and Zisserman[19] use a clustering method on transaction data to produce typical prototypes of visual patterns.

To reduce the quantization error of visual primitives and eliminate the ambiguities among visual patterns, Yuan and Wu[33] propose the context-aware clustering algorithm. In their work, the disambiguation of visual words and the discovery of visual patterns are optimized by a self-supervised clustering procedure that allows visual feature quantization and visual pattern clustering to help each other, thus leading to a better visual vocabulary as well as better visual patterns. Further, Wang et al.[54] extend the context-aware clustering method by incorporating multiple types of features. Their work provides a uniform solution that can handle visual patterns in both spatial and feature spaces.

Most abovementioned methods ignore the frequencies of primitive occurrence in the local spatial neighborhood. Kim et al.[47] thus propose the bag-to-set (B2S) approach to encode visual word frequencies occurring in each local spatial neighborhood into a long binary vector, which is used for visual pattern mining. However, this method tends to generate artificial visual patterns not present in given datasets. An alternative approach proposed by Fernando et al.[68] exploits the frequency information of visual words during the process of discriminative visual pattern mining. This method effectively avoids the generation of artificial visual patterns that may cause performance loss. Besides, Kim et al.[34] allow replicated visual words to appear in a bag instead of containing distinct visual words in a set and propose a spatial item bag mining method, which finds frequent visual patterns according to semi-affine invariance of spatial layout among objects in image data. Furthermore, a spatial relationship pattern-based hierarchical clustering algorithm is developed to cluster those similar object patterns.

Visual Co-occurrence Matching and Counting for Visual Pattern Discovery

To apply classic FIM methods, one needs to build transactions based on the visual vocabulary of image or video data in advance. The discovery of visual patterns will heavily depend on the quality of transactions, and further depend on the quality of the visual vocabulary. These dependencies can be mitigated by frequent pattern counting methods as they do not need build transaction data. For example, Zhang and Chen[40] propose to mine visual patterns in offset space, which is extended by Zhang et al.[48,57] and Li et al.[8] An offset space is generated by the relative location difference of visual primitives between two images in

TABLE 1 | Representative Work of Visual Pattern Discovery

Authors and References	Year	Method	Taxonomy	Data source	Dataset
Hong and Huang[16]	2000	Maximum likelihood method	Bottom-up	Image	Proprietary
Hsu et al.[17]	2003	Viewpoint mining	Bottom-up	Image set	Proprietary
Fergus et al.[18]	2003	Constellation model	Top-down	Image set	Caltech 4, UIUC car, Corel
Sivic and Zisserman[19]	2004	Feature neighborhood clustering	Bottom-up	Video	Groundhog Day, Fawlty Towers
Hong and Huang[20]	2004	Probabilistic parametric model	Top-down	Image set	Proprietary
Leordeanu and Hebert[21]	2005	Subgraph mining	Bottom-up	Image set	Proprietary
Felzenszwalb and Huttenlocher[22]	2005	Pictorial model	Top-down	Image set	Proprietary
Zhu et al.[6]	2005	Texton learning	Top-down	Image, video	Proprietary
Sivic et al.[23]	2005	Probabilistic Latent semantic analysis (pLSA)	Top-down	Image set	Caltech 101, MIT indoor/outdoor
Fei-Fei and Perona[24]	2005	Bayesian hierarchical model	Top-down	Image set	Scene 13
Wang et al.[25]	2006	Dependent regions	Top-down	Image set	Caltech 4, Caltech 101
Russell et al.[26]	2006	Latent Dirichlet allocation (LDA)	Top-down	Image set	Caltech 4, MSRC v2, LabelMe
Quack et al.[27]	2007	Spatial association rule mining	Bottom-up	Image set	ETHZ, GRAZ, TUD, CALTECH
Yuan and Wu[28]	2007	Spatial random partition	Bottom-up	Image set	Proprietary
Lee et al.[29]	2007	Spatial association rule mining	Bottom-up	Image set, video	Proprietary
Fidler and Leonardis[7]	2007	Hierarchical part composition learning	Bottom-up	Image set	Proprietary, UIUC car
Cao and Li[30]	2007	Spatially coherent latent topic model	Top-down	Image set	Weizmann horses, LOCUS horses, Microsoft cow, Scene 13, Caltech 101
Liu and Chen[31]	2007	Spatial-temporal model	Top-down	Video	Helicopter sequence, Car sequence
Gilbert et al.[32]	2008	Compound features mining	Bottom-up	Video	KTH
Yuan and Wu[33]	2008	Context-aware clustering	Bottom-up	Image	PSU Near-Regular Texture
Kim et al.[34]	2008	Spatial item bag mining	Bottom-up	Image set	Proprietary
Liu et al.[35]	2008	Spatial histograms	Bottom-up	Image set	PASCAL VOC06, Caltech 4, MSRC v2
Todorovic and Ahuja[36]	2008	Tree structural subimage matching	Top-down	Image set	Caltech 101, Caltech rear-view cars, UIUC multiscale side-view cars, Weizmann side-view horses, TUD side-view cows
Sivic et al.[37]	2008	Hierarchical latent Dirichlet allocation(hLDA)	Top-down	Image set	MSRC B1
Tang and Lewis[38]	2008	Non-negative matrix factorization (NMF)	Top-down	Image set	Washington images, LabelMe
Gao et al.[39]	2009	Frequent subgraph mining	Bottom-up	Image, image set	Proprietary, Caltech 101
Zhang and Chen[40]	2009	Clustering in offset space	Bottom-up	Image set	Caltech 101, MSRC v2, Graz 01

TABLE 1 | Continued

Authors and References	Year	Method	Taxonomy	Data source	Dataset
Lee and Grauman[41]	2009	Spectral graph clustering	Bottom-up	Image set	Caltech 101, ETHZ shape, LabelMe
Payet and Todorovic[42]	2009	Coordinate ascent Swendsen-Wang cut	Bottom-up	Image set	Caltech 101, ETHZ shape, LabelMe, Weizmann horses
Zheng et al.[43]	2009	Visual synset	Bottom-up	Image set	Caltech 256
Chum et al.[44]	2009	Geometric min-hash	Bottom-up	Image set	Oxford buildings 5K
Tan and Ngo[45]	2009	Localized matching using Earth Mover's Distance	Bottom-up	Image set	Proprietary
Endres et al.[46]	2009	Latent Dirichlet allocation (LDA)	Top-down	Range image	Proprietary
Kim et al.[47]	2010	Frequent item bag mining	Bottom-up	Image set	Caltech 101
Zhang and Chen[48]	2010	Voting in offset space	Bottom-up	Image set	Pascal VOC05, Graz 01, Graz 02, Caltech 4
Heath et al.[49]	2010	Affine cosegmentation	Bottom-up	Image set	Proprietary
Liu and Yan[50]	2010	Subgraph mining	Bottom-up	Image set	Columbia near duplicate images, IST faces
Bagon et al.[51]	2010	Ensemble matching	Bottom-up	Image set	ETHZ shapes
Cho et al.[52]	2010	Multilayer match-growing	Bottom-up	Image set	ETHZ shapes, Proprietary
Liu et al.[53]	2010	Hierarchical visual model	Top-down	Video	Proprietary, TRECVID
Wang et al.[54]	2011	Multicontext-aware clustering	Bottom-up	Image, image set	Proprietary, MSRC v2
Zhao et al.[55]	2011	Cohesive sub-graph mining	Bottom-up	Video	Proprietary
Wang et al.[56]	2011	Emerging pattern mining	Bottom-up	Video	KTH, YouTube, Proprietary
Zhang et al.[57]	2011	Voting in offset space	Bottom-up	Image set	Oxford buildings 5K, Flicker 1M
Zhang et al.[58]	2011	Descriptive visual phrases	Bottom-up	Image set	Proprietary, Corel 5k, Caltech 101, Caltech 256
Zhang et al.[59]	2011	Contextual visual vocabulary	Bottom-up	Image set	Proprietary
Sun and Hamme[60]	2011	Graph regularized NMF	Top-down	Image set	Caltech 256
Philbin et al.[61]	2011	Geometric Latent Dirichlet allocation (gLDA)	Top-down	Image set	Oxford buildings 5K, Statue of Liberty 37K, Rome 1K
Sadeghi and Farhadi[62]	2011	Max margin structure learning	Top-down	Image set	UIUC Phrase
Yuan and Wu[63]	2012	Self-supervised subspace learning	Bottom-up	Image set	Caltech 101
Yuan et al.[64]	2012	Multilayer candidate pruning	Bottom-up	Image set, video	Proprietary
Wang et al.[65]	2012	Actionlet ensemble mining	Bottom-up	Video	MSR Action3D, MSR Daily Activity3D
Lee and Grauman[66]	2012	Object-graphs	Bottom-up	Image set	MSRC v0, MSRC v2, Corel, PASCAL VOC08, Gould 2009
Li et al.[8]	2012	Voting in offset space	Bottom-up	Image set	UIUC phrase, PASCAL VOC07, SUN 09, MIT indoor
Faktor and Irani[67]	2012	Clustering by composition	Bottom-up	Image set	Caltech 101, ETHZ shape, Pascal VOC 2010, Ballet-Yoga
Fernando et al.[68]	2012	Frequent local histogram mining	Bottom-up	Image set	GRAZ 01, Oxford flower 17, Scene 15, PASCAL VOC07

TABLE 1 | Continued

Authors and References	Year	Method	Taxonomy	Data source	Dataset
Singh et al.[69]	2012	Discriminative doublets	Bottom-up	Image set	MIT indoor 67
Jiang et al.[70]	2012	Randomized visual phrases	Bottom-up	Image set	Groundhog Day, Belgalogo, Proprietary
Hao et al.[71]	2012	3D visual phrases	Bottom-up	Image set	Proprietary, Oxford buildings 5K
Chu and Tsai[72]	2012	Frequent subgraph mining	Bottom-up	Image set	Proprietary, Oxford Paris
Zhu et al.[73]	2012	Saliency-guided multiple class learning	Bottom-up	Image set	SIVAL, iCoseg, 3D object category
Cong et al.[74]	2012	Sparse dictionary selection	Bottom-up	Video	Kodak Home Video
Zhang and Tao[75]	2012	Slow feature analysis	Bottom-up	Video	KTH, Weizmann, CASIA Interaction, UT Interaction
Niu et al.[76]	2012	Context aware topic model	Top-down	Image set	Scene 15, LabelMe, UIUC sports
Andreetto et al.[77]	2012	Affinity-based LDA model (A-LDA)	Top-down	Image set	Egrets 100, MSRC v1, MSRC v2, Scene 8, LabelMe
Cong et al.[78]	2013	Sparse reconstruction cost	Bottom-up	Video	UMN, UCSD, Subway
Rubinstein et al.[79]	2013	Reliable matching and saliency detection	Bottom-up	Image set	MSRC, iCoseg, Proprietary
Song et al.[80]	2013	Hierarchical sequence summarization	Bottom-up	Video	Arm Gesture, Canal 9, NATOPS
Wang et al.[81]	2013	Spatiotemporal part sets mining	Bottom-up	Video	UCF sport, Keck gesture, MSR-Action3D
Li et al.[82]	2013	Mid-level visual concept learning	Bottom-up	Image set	PASCAL VOC07, Scene 15, MIT indoor, UIUC sports, Inria horse
Myeong and Lee[83]	2013	High-order semantic relation transfer	Bottom-up	Image set	LableMe 19-class, LabelMe outdoor, Polo
Zhao et al.[84]	2013	LDA with word co-occurrence prior	Top-down	Video	Proprietary

For papers that have both conference and journal versions, only journal versions are listed.

the sense of scale alignment. Such offset space enables co-occurring visual primitives to be assembled into the near-same place, thus facilitating visual pattern discovery. In Ref 40, the visual primitives having the absolutely same location in the offset space compose a high-order visual pattern. Allowing slight deformation, Hough voting is further adopted to highlight the frequent co-occurring visual primitives in Refs 8, 48 and 57. It is worth pointing out that those studies in Refs 40, 48 and 57 focus on mining compositional patterns of image feature patches, while Ref 8 is engaged in automatic discovery of object group patterns.

Primitive matching and counting can be used in many other ways for common pattern discovery. Hong and Huang[16] apply template matching and maximum likelihood criteria for common object discovery. Yuan and Wu[28] introduce the spatial random partition method, which randomly partitions each image several times to generate a pool of subregions. Then common objects can be discovered by finding frequent feature matches in the subregion pool. Bagon et al.[51] detect and sketch common objects from multiple images by candidate region matching. Cho et al.[52] present a multilayer match-growing method to discover common objects from a single or multiple images. Zhao and Yuan[89] and Yuan et al.[64] propose the multilayer candidate pruning approach for common object discovery, where set-to-set matching and branch-and-bound search are applied. Fidler and Leonardis[7] learn a hierarchical representation for each object category by feature indexing and matching. Liu and Liu[90] find optimal visual word matches and discover common object patterns by a greedy randomized adaptive search procedure. Faktor and Irani[67] develop the 'clustering by composition' method for common scene pattern discovery. With the assumption that images from the same class can generate each other by shared regions, they discover image class patterns by a collaborative randomized matching region search algorithm.

Graph-Based Mining for Visual Pattern Discovery

Since a graph can directly model sophisticated spatial structures among visual primitives, many approaches have been proposed to discover visual patterns using graph mining rather than FIM and visual co-occurrence matching/counting. A typical case is that Gao et al.[39] use a frequent subgraph pattern mining method to discover high-order geometric visual patterns. They encode the spatial relationship of each pair of visual words into a link vector. The pairwise visual word associations can then be identified according to the spatial consistent rules. Using frequent subgraph pattern mining on the association graph, the high-order geometric patterns can be obtained. Owing to the invariant representation of the spatial relationship between each pair of visual words, the obtained high-order patterns also exhibit translation, scale, and rotation invariance.

Besides mining visual patterns with a fixed order, graph mining is also used to extract common visual patterns without order constraint. For instances, Leordeanu and Hebert[21] and Liu and Yan[50] employ subgraph mining on a feature correspondence graph of two images to discover common image patterns. Recently, Zhao et al.[55,5] have also proposed a cohesive subgraph mining method to find thematic patterns in video, where the overall mutual information scores among the spatiotemporal visual words are maximized. To model common object shapes, Lee and Grauman[41] perform matching on patch-anchored edge fragments, and spectral graph clustering is performed for common shape discovery. Similarly, Payet and Todorovic[42] build graph on all pairs of geometric matched contours in images to discover common object contours.

TOP-DOWN PATTERN MINING

The bottom-up pattern discovery method starts with unordered visual primitives and then merges the primitives until larger visual patterns are found. In contrast, the top-down method starts with the modeling of visual patterns and then infers the pattern discovery result.

Inspired by the success of unsupervised topic discovery in statistical natural language processing, most of top-down methods use generative topic models for visual pattern modeling.[23,25,26,37] In this section, we first review the classic topic model based visual pattern discovery methods. Then we turn to the methods that incorporate spatial and temporal constraints into topic models.[20,31,36,61,84,91] After that, we discuss subspace projection methods for visual pattern discovery.[38,60]

Classic Topic Model for Visual Pattern Discovery

The topic model, such as LDA[15] and probabilistic latent semantic analysis (pLSA),[92] discovers semantic topics from a corpus of documents. Generally, the 'bag-of-words' representation is used to model the documents. Meanwhile, each word is generated from one topic while each document is modeled as a probability distribution of the latent topics.

Sivic et al.[23] use the topic model to discover and locate objects in images. They use the bag-of-words model to represent each image and consider the local co-occurring regions by the ''doublets'' pairs of visual words. This model treats each image as a histogram of visual words. After obtaining all documents, the pLSA model is used to discover the object topics. This method can discover the object categories and localize the object instances in the image. Following this idea, Russell et al.[26] discover the visual object categories based on the LDA and pLSA model. To group visual words spatially, they first segment the images multiple times and then discover object topics from a pool of segments. The discovered topics are closely related to the ground-truth object classes. To discover the hierarchical structure for the visual patterns, Sivic et al.[37] investigate the hierarchical LDA (hLDA) model. Based on the multiple segmentation framework,[26] this method can automatically discover the object hierarchies from image collections.

Besides using the segmentation of each single image as Ref 26, Andreetto et al.[77] combine a LDA model and a hybrid parametric–nonparametric model for categorical object discovery and segmentation. This method segments multiple images simultaneously while the segments in different images benefit from each other. By sharing the shape and appearance information of each segment, it can improve the object discovery and segmentation performance simultaneously.

Topic Model with Spatial and Temporal Constraints for Visual Pattern Discovery

Besides the frequency of visual features captured by 'bag-of-words' representation, the spatial and temporal contexts are also important cues for visual pattern modeling. To better encode spatial structures among visual words, Wang and Grimson[91] propose a spatial LDA (sLDA) model. The word-document assignment is no longer a fixed prior, but varies depending on a generative procedure, in which visual words will be assigned into the same document if they are close in image space. Philbin et al.[61] introduce the geometric LDA (gLDA) model for object discovery from a corpus of images. As an extension of LDA, gLDA

considers the affine homographic geometric relation in the generative process. The gLDA model has better performance than the standard LDA model in the application of particular object discovery. Besides encoding the two-dimensional spatial structures in the image, Endres et al.[46] apply LDA model to discover objects in 3D range data directly.

Liu and Chen[31] extend topic models from still images to videos with a temporal model integrated. The topic model is used for appearance modeling while the probabilistic data association (PDA) filter is used for motion modeling. By tightly integrating the spatial and temporal models, they show promising video object discovery results. To engage human in the loop for video object discovery, Liu et al.[53] employ the topic model in a semi-supervised learning framework. By taking weakly supervised information from human, their model can be tailored to users' interests for targeted object discovery.

Recently, Zhao et al.[84] notice that important co-occurrence information among local features is ignored in the LDA model. To tackle this issue, they propose to incorporate a Gaussian Markov word co-occurrence prior into the general LDA model, such that bottom-up induction and top-down deduction can help each other for efficient topic video object discovery.

Besides introducing spatial constraints into pLSA/LDA models, there are also methods that explicitly use graph or tree to model the spatial structure of visual patterns, e.g., Refs 36 and 20. Unlike pLSA/LDA based methods, Hong and Huang[20] model the visual pattern as a mixture of probabilistic parametric attributed relational graphs while each image is represented by an attributed relational graph in image (spatial) space. They also propose an expectation-maximization (EM) algorithm to learn the parameters of visual pattern model. In addition, Todorovic and Ahuja's[36] method is also different from pLSA/LDA-based methods, which models the spatial layout of primitive regions in a tree structure to learn common object category.

Subspace Projection for Visual Pattern Discovery

Apart from the statistical viewpoint to mine visual patterns, e.g., pLSA and LDA based model, there are also subspace projection methods to approximate the semantic structure of visual patterns. A typical approach is to perform non-negative matrix factorization (NMF). For the detailed discussion about the equivalence between pLSA and NMF, refer Refs 93 and 94. In terms of visual pattern discovery using NMF, Tang and Lewis's[38] work is a good practice. They show that the results of NMF are comparable with that of LDA on the same dataset. It is also worth mentioning that Sun and Hamme[60] incorporate NMF to model recurring visual patterns and spectral clustering to cluster visual primitives into visual patterns.

SUMMARY OF BOTTOM-UP AND TOP-DOWN METHODS

Bottom-up methods proceed from the local layout of visual primitives to recognize general visual patterns in image and video data. Such methods emphasize on assembling visual primitives into visual patterns. There are several advantages of bottom-up methods. First of all, bottom-up methods can be widely applied for their data-driven nature. Second, bottom-up methods can easily incorporate varieties of contexts such as spatial co-occurrence of multiple visual primitives and geometric relationship between pairs of visual primitives. Third, bottom-up methods are easy to implement. However, bottom-up methods mainly investigate local spatial cues of visual patterns while lack global modeling of visual patterns.

In contrast, top-down methods work the other way around, which treat images or videos as mixture patterns over visual primitives in a global perspective. Such methods focus on modeling and inferring the composition of visual patterns. There are several advantages of top-down methods. First of all, the top-down methods can deal with variations of visual patterns by using probability reasoning for their modeling of visual data. Second, the top-down methods can discover multiple visual patterns simultaneously as the generative model is naturally designed for modeling multiple patterns. Third, the top-down methods can also incorporate the spatial and geometry information of visual patterns. However, it is not trivial to handle model parameter learning and posterior probability inference for top-down methods.

Choosing between bottom-up and top-down approaches is application – dependent. Generally, when we observe a number of specific spatial compositions of visual primitives and expect from them to infer common visual patterns, bottom-up methods will be appropriate; while when we are required to model pattern mixture and reason posterior distribution of visual pattern mixture over visual primitives, top-down methods should be preferable.

APPLICATIONS

Visual patterns capture the spatial layout of visual primitives, e.g., local features, segments, objects.

Such meaningful patterns can contribute to many applications, such as image search,[43,45,49,57,59,70,72] object categorization,[25,35,58,63,66,73,79,82] scene recognition,[8,24,30,62,69,71,76,83] and video analysis.[19,32,53,55,56,64,65,75,80,81]

Image Search

Visual patterns offer information-rich visual phrase retrieval compared to image retrieval using bag-of-visual-word representation. Several approaches have been proposed, including visual synset,[43] geometry preserving visual phrases,[57] contextual visual vocabulary,[59] and randomized visual phrases.[70] In Ref 43, a higher-level visual representation derived from visual word patterns, visual synset, is proposed by Zheng et al. to improve the performance of image retrieval. In addition to exploring visual word co-occurrences, the visual phrases proposed by Zhang et al.[57] also capture the geometric relationships among visual words, thus present a better retrieval performance than traditional bag-of-visual-words model. To better retrieve near-duplicate images, a spatial contextual visual vocabulary method considering local feature group is proposed by Zhang et al.[59] Combining with spatial random partition,[28] randomized visual phrases are constructed by Jiang et al.[70] for more discriminative matching in visual object search methods. Besides constructing visual phrase descriptors for image retrieval, there are also pattern matching based methods[45,49,72] and min-hashing scheme.[44] Tan and Ngo[45] utilize localized matching for query-by-pattern image search. Heath et al.[49] perform image search by connectivity among visual patterns in images. Chu and Tsai[72] perform product image search by motif pattern matching. Chum et al.[44] propose geometric min-hashing index for object discovery and image retrieval.

Object Categorization

Visual patterns are also beneficial to object categorization. In Ref 63, Yuan and Wu leverage the discovered visual phrase lexicon obtained by FIM and subspace learning to effectively reduce the ambiguity between foreground and background objects. In the work by Zhang et al.,[58] the frequent occurring visual word pairs are used to construct the descriptive visual phrases for an effective representation of certain visual objects. Owing to the consideration of co-occurrences of image patches, the method proposed by Wang et al.[25] shows high competitive object categorization ability. In Ref 35, Liu et al. integrate feature selection and higher-order spatial feature extraction together for an efficient object categorization. The method proposed by Lee and Grauman[66] leverages object co-occurrence patterns for visual object categorization. Zhu et al.[73] use saliency-guided multiple class learning to discover object patterns and perform object categorization. Rubinstein et al.[79] separate the common category of objects from noisy image collections by reliable matching and saliency detection. The mid-level visual concepts are exploited by Li et al.[82] to harvest visual patterns from images and help enhance the object classification performance.

Scene Recognition

Scene recognition is another application of visual patterns. The spatial co-occurrences of image patches are used for a better scene representation by Singh et al.[69] The method proposed by Hao et al.[71] constructs 3D visual phrases with particular geometric structures for landmark recognition. Niu et al.[76] leverage the spatial layout of image patches to design a context-aware topic model for scene recognition. Bayesian hierarchical model proposed by Fei-Fei and Perona[24] and spatially coherent latent topic model proposed by Cao and Li[30] also describe visual patterns as a topic model for scene recognition. The proposed visual phrase detector by Sadeghi and Farhadi[62] encodes the interaction between objects or activities of single objects for phrasal recognition and object detection. Li et al.[8] make use of the recurring compositions of objects across images for a better scene categorization. Myeong and Lee[83] perform label transfer on high-order relations of objects for scene segmentation and semantic region recognition.

Video Analysis

Video analysis also needs the effective extraction of visual patterns. In Sivic and Zisserman's work,[19] the spatial configurations of viewpoint invariant features are mined for movie summarization. In Ref 5, Zhao et al. extract key action patterns for sports video summarization. Liu et al.,[53] Zhao et al.,[55] and Yuan et al.[64] discover thematic patterns to highlight products appearing in commercial advertisements and perform video object summarization. Cong et al.[74] utilize the sparsity consistency of visual patterns to construct a sparse representative dictionary towards video summarization.

Besides video summarization, visual patterns can be used for video anomaly detection. For example, by mining the normal event patterns, one can identify the rest as anomalies. In Ref 95, Jiang et al. discover regular rules of normal events from spatiotemporal context and perform video anomaly detection. Cong et al.[78,96] apply sparse reconstruction over the

normal motion patterns to detect abnormal events in videos.

Another typical application of visual patterns in video analysis is to recognize human actions. Gilbert et al.[32] identify compound patterns from frequently co-occurring dense spatiotemporal corners for action recognition. Wang et al.[56] learn discriminative features by mining emerging patterns[97] for instant action recognition in a video. Wang et al.[65] represent a particular conjunctive pattern of joint locations as actionlet and recognize actions by actionlet ensemble structure mining. Zhang and Tao[75] extract useful motion patterns from videos for human action recognition by slow feature analysis (SFA).[98] Song et al.[80] perform action recognition by grouping similar spatiotemporal patterns in a hierarchical sequence summarization framework. Wang et al.[81] recognize human actions by mining distinctive co-occurring spatial configurations of body parts in spatial domain and pose movement patterns in temporal domain.

CONCLUSION AND OUTLOOK

Over the past decade, visual pattern discovery has received increasing attention, especially by the communities of computer vision and data mining. In this survey, we have collected the abundant literature of visual pattern discovery, and discussed both bottom-up and top-down techniques as well as their diverse applications. In the bottom-up methods, the common strategy is to mine visual co-occurrence compositions from local neighborhoods of visual primitives (e.g., local image patches, segments, objects). The top-down methods are usually built on varieties of topic models, which are used to infer the pattern discovery result for either image or video data.

Although tremendous progress has been made, there are still several open issues that need to be addressed in future work, including: (1) how to interpret visual patterns and effectively measure their quality; (2) how to select representative and discriminative patterns; (3) how to suitably integrate multiple complementary feature modalities for visual pattern discovery; and (4) how to effectively combine the bottom-up and top-down approaches of visual pattern discovery.

Firstly, the interpretation and quality measure is crucial to visual pattern discovery. Despite a few successes in explaining visual patterns,[6,67,8] we still need deeper investigation of spatial co-occurrences, geometric associations, and visual appearance of individual primitives, in order to better understand and utilize visual patterns.

Secondly, mining representative and discriminative patterns is a nontrivial problem as sometimes the two goals contradict to each other. However, depending on application, it is interesting to develop methods that can find such visual patterns, e.g., local frequent histograms[68] and discriminative doublets.[69]

Thirdly, image and video data naturally exhibit multiple feature modalities that are complementary. Most existing approaches discover visual patterns using a single feature modality. However, for a better visual pattern discovery, a suitable integration of multiple complementary features ought to be studied.[54]

Finally, bottom-up methods capture local spatial cues of visual patterns while top-down methods model compositions of visual patterns.[60,84] How to combine the strengths of bottom-up methods and top-down methods for visual pattern discovery is an interesting research topic.

ACKNOWLEDGMENT

This work is supported in part by the Nanyang Assistant Professorship SUG M4080134.

REFERENCES

1. Tuytelaars T, Lampert C, Blaschko M, Buntine W. Unsupervised object discovery: a comparison. *Int J Comput Vis* 2010, 88:284–302.
2. Yuan J. Discovering visual patterns in image and video data: concepts, algorithms, experiments. Saarbrücken, Germany: VDM Verlag Dr. Müller 2011.
3. Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: *CVPR Workshop on Generative-Model Based Vision*, Washington, DC; 2004, 178–178.
4. Quattoni A, Torralba A. Recognizing indoor scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL; 2009.
5. Zhao G, Yuan J. Discovering thematic patterns in videos via cohesive sub-graph mining. In: *Proceedings*

of the IEEE International Conference on Data Mining, Vancouver, BC; 2011, 1260–1265.

6. Zhu S, Guo C, Wang Y, Xu Z. What are textons? *Int J Comput Vis* 2005, 62:121–143.
7. Fidler S, Leonardis A. Towards scalable representations of object categories: Learning a hierarchy of parts. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN; 2007, 1–8.
8. Li C, Parikh D, Chen T. Automatic discovery of groups of objects for scene understanding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI; 2012, 2735–2742.
9. Thompson DW. *On growth and Form*. Cambridge, UK: Cambridge University Press; 1961.
10. Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. *Data Mining Knowl Discov* 2007, 15:55–86.
11. Tuytelaars T, Mikolajczyk K. Local invariant feature detectors: a survey. *Found Trends Comput Graph Vis* 2008, 3:177–280.
12. Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2000, 22:888–905.
13. Felzenszwalb P, Girshick R, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Trans Patt Anal Mach Intell* 2010, 32:1627–1645.
14. Grauman K, Leibe B. *Visual Object Recognition (Synthesis Lectures on Artificial Intelligence and Machine Learning)*. San Rafael, CA: Morgan & Claypool Publishers; 2011.
15. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003, 3:993–1022.
16. Hong P, Huang T. Extracting the recurring patterns from image. In: *Proceedings of the Asian Conference on Computer Vision*, Taipei, Taiwan; 2000, 8–11.
17. Hsu W, Dai J, Lee M. Mining viewpoint patterns in image databases. In: *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington, DC; 2003, 553–558.
18. Fergus R, Perona P, Zisserman A. Object class recognition by unsupervised scale-invariant learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *Volume 2*, Dublin, Ireland; 2003, II–264.
19. Sivic J, Zisserman A. Video data mining using configurations of viewpoint invariant regions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC; 2004, I–488.
20. Hong P, Huang T. Spatial pattern discovery by learning a probabilistic parametric model from multiple attributed relational graphs. *Discrete Appl Math* 2004, 139:113–135.
21. Leordeanu M, Hebert M. A spectral technique for correspondence problems using pairwise constraints. In: *Proceedings of the IEEE International Conference on Computer Vision*, *Volume 2*, Beijing, China; 2005, 1482–1489.
22. Felzenszwalb P, Huttenlocher D. Pictorial structures for object recognition. *Int J Comput Vis* 2005, 61:55–79.
23. Sivic J, Russell B, Efros A, Zisserman A, Freeman W. Discovering objects and their location in images. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2005, 370–377.
24. Fei-Fei L, Perona P. A Bayesian hierarchical model for learning natural scene categories. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Beijing, China; 2005, 524–531.
25. Wang G, Zhang Y, Fei-Fei L. Using dependent regions for object categorization in a generative framework. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York; 2006, 1597–1604.
26. Russell B, Freeman W, Efros A, Sivic J, Zisserman A. Using multiple segmentations to discover objects and their extent in image collections. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York; 2006, 1605–1614.
27. Quack T, Ferrari V, Leibe B, Van Gool L. Efficient mining of frequent and distinctive feature configurations. In: *Proceedings of the IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil; 2007.
28. Yuan J, Wu Y. Spatial random partition for common visual pattern discovery. In: *Proceedings of the IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil; 2007, 1–8.
29. Lee A, Hong R, Ko W, Tsao W, Lin H. Mining spatial association rules in image databases. *Inform Sci* 2007, 177:1593–1608.
30. Cao L, Fei-Fei L. Spatially coherent latent topic model for concurrent object segmentation and classification. In: *Proceedings of the IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil; 2007.
31. Liu D, Chen T. A topic-motion model for unsupervised video object discovery. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN; 2007.
32. Gilbert A, Illingworth J, Bowden R. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In: *Proceedings of the European Conference on Computer Vision*, Marseille, France; 2008, 222–233.
33. Yuan J, Wu Y. Context-aware clustering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK; 2008, 1–8.
34. Kim S, Jin X, Han J. Sparclus: spatial relationship pattern-based hierarchical clustering. In *Proceedings of*

the SIAM International Conference on Data Mining, Atlanta, GA; 2008.

35. Liu D, Hua G, Viola P, Chen T. Integrated feature selection and higher-order spatial feature extraction for object categorization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK; 2008, 1–8.
36. Todorovic S, Ahuja N. Unsupervised category modeling, recognition, and segmentation in images. *IEEE Trans Pattern Anal Mach Intell* 2008, 30:2158–2174.
37. Sivic J, Russell B, Zisserman A, Freeman W, Efros A. Unsupervised discovery of visual object class hierarchies. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK; 2008, 1–8.
38. Tang J, Lewis PH. Non-negative matrix factorisation for object class discovery and image auto-annotation. In: *Proceedings of the International Conference on Content-based Image and Video Retrieval*, Niagara Falls, Canada; 2008, 105–112.
39. Gao J, Hu Y, Liu J, Yang R. Unsupervised learning of high-order structural semantics from images. In: *Proceedings of the IEEE International Conference on Computer Vision*, Kyoto, Japan; 2009, 2122–2129.
40. Zhang Y, Chen T. Efficient kernels for identifying unbounded-order spatial features. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL; 2009, 1762–1769.
41. Lee Y, Grauman K. Shape discovery from unlabeled image collections. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL; 2009, 2254–2261.
42. Payet N, Todorovic S. From a set of shapes to object discovery. In: *Proceedings of the European Conference on Computer Vision*, Heraklion, Crete; 2010, 57–70.
43. Zheng Y, Neo S, Chua T, Tian Q. Toward a higher-level visual representation for object-based image retrieval. *Vis Comput* 2009, 25:13–23.
44. Chum O, Perdoch M, Matas J. Geometric min-hashing: Finding a (thick) needle in a haystack. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL; 2009, 17–24.
45. Tan H, Ngo C. Localized matching using Earth Mover'sDistance towards discovery of common patterns from small image samples. *Image Vis Comput* 2009, 27:1470–1483.
46. Endres F, Plagemann C, Stachniss C, Burgard W. Unsupervised discovery of object classes from range data using latent Dirichlet allocation. In: *Robotics: Science and Systems*, Seattle, Washington; 2009.
47. Kim S, Jin X, Han J. Disiclass: discriminative frequent pattern-based image classification. In: *KDD Workshop on Multimedia Data Mining*, Washington, DC; 2010, 7.
48. Zhang Y, Chen T. Weakly supervised object recognition and localization with invariant high order features. In: *Proceedings of the British Machine Vision Conference*, Aberystwyth, UK; 2010, 47.
49. Heath K, Gelfand N, Ovsjanikov M, Aanjaneya M, Guibas L. Image webs: computing and exploiting connectivity in image collections. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI; 2010, 3432–3439.
50. Liu H, Yan S. Common visual pattern discovery via spatially coherent correspondences. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA; 2010:1609–1616.
51. Bagon S, Brostovski O, Galun M, Irani M. Detecting and sketching the common. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA; 2010, 33–40.
52. Cho M, Shin YM, Lee KM. Unsupervised detection and segmentation of identical objects. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA; 2010, 1617–1624.
53. Liu D, Hua G, Chen T. A hierarchical visual model for video object summarization. *IEEE Trans Pattern Anal Mach Intell* 2010, 32:2178–2190.
54. Wang H, Yuan J, Tan Y. Combining feature context and spatial context for image rattern discovery. In: *Proceedings of the IEEE International Conference on Data Mining*, Vancouver, Canada; 2011, 764–773.
55. Zhao G, Yuan J, Xu J, Wu Y. Discovery of the thematic object in commercial videos. *IEEE Multimedia Mag* 2011, 18:56–65.
56. Wang L, Wang Y, Jiang T, Gao W. Instantly telling what happens in a video sequence using simple features. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO; 2011, 3257–3264.
57. Zhang Y, Jia Z, Chen T. Image retrieval with geometry-preserving visual phrases. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO; 2011, 809–816.
58. Zhang S, Tian Q, Hua G, Huang Q, Gao W. Generating descriptive visual words and visual phrases for large-scale image applications. *IEEE Trans Image Process* 2011, 20:2664–2677.
59. Zhang S, Tian Q, Hua G, Zhou W, Huang Q, Li H, Gao W. Modeling spatial and semantic cues for large-scale near-duplicated image retrieval. *Comput Vis Image Understand* 2011, 115:403–414.
60. Sun M, Hamme HV. Image pattern discovery by using the spatial closeness of visual code words. In: *Proceedings of the IEEE International Conference on Image Processing*, Brussels, Belgium; 2011, 205–208.
61. Philbin J, Sivic J, Zisserman A. Geometric latent Dirichlet allocation on a matching graph for large-scale image datasets. *Int. J. Comput. Vis.* 2011, 95:138–153.

62. Sadeghi M, Farhadi A. Recognition using visual phrases. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO; 2011, 1745–1752.
63. Yuan J, Wu Y. Mining visual collocation patterns via self-supervised subspace learning. *IEEE Trans Syst Man Cybern B Cybern* 2012, 42:1–13.
64. Yuan J, Zhao G, Fu Y, Li Z, Katsaggelos A, Wu Y. Discovering thematic objects in image collections and videos. *IEEE Trans Image Process* 2012, 21:2207–2219.
65. Wang J, Liu Z, Wu Y, Yuan J. Mining actionlet ensemble for action recognition with depth cameras. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI; 2012, 1290–1297.
66. Lee Y, Grauman K. Object-graphs for context-aware visual category discovery. *IEEE Trans Pattern Anal Mach Intell* 2012, 34:346–358.
67. Faktor A, Irani M. ''Clustering by composition''—Unsupervised discovery of image categories. In: *Proceedings of the European Conference on Computer Vision*, Florence, Italy; 2012, 474–487.
68. Fernando B, Fromont E, Tuytelaars T. Effective use of frequent itemset mining for image classification. In: *Proceedings of the European Conference on Computer Vision*, Florence, Italy, 2012.
69. Singh S, Gupta A, Efros A. Unsupervised discovery of mid-level discriminative patches. In: *Proceedings of the European Conference on Computer Vision*, Firenze, Italy; 2012.
70. Jiang Y, Meng J, Yuan J. Randomized visual phrases for object search. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI; 2012, 3100–3107.
71. Hao Q, Cai R, Li Z, Zhang L, Pang Y, Wu F. 3D visual phrases for landmark recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI; 2012, 3594–3601.
72. Chu WT, Tsai MH. Visual pattern discovery for architecture image classification and product image search. In: *Proceedings of the ACM International Conference on Multimedia Retrieval*, Hong Kong, China; 2012, 27:1–27:8.
73. Zhu JY, Wu J, Wei Y, Chang E, Tu Z. Unsupervised object class discovery via saliency-guided multiple class learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI; 2012, 3218–3225.
74. Cong Y, Yuan J, Luo J. Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE Trans Multimedia* 2012, 14:66–75.
75. Zhang Z, Tao D. Slow feature analysis for human action recognition. *IEEE Trans Pattern Anal Mach Intell* 2012, 34:436–450.
76. Niu Z, Hua G, Gao X, Tian Q. Context aware topic model for scene recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI; 2012, 2743–2750.
77. Andreetto M, Zelnik-Manor L, Perona P. Unsupervised learning of categorical segments in image collections. *IEEE Trans Patt Anal Mach Intell* 2012, 34:1842–1855.
78. Cong Y, Yuan J, Liu J. Abnormal event detection in crowded scenes using sparse representation. *Patt Recogn* 2013, 46:1851–1864.
79. Rubinstein M, Joulin A, Kopf J, Liu C. Unsupervised joint object discovery and segmentation in Internet images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR; 2013.
80. Song Y, Morency LP, Davis R. Action recognition by hierarchical sequence summarization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR; 2013.
81. Wang C, Wang Y, Yuille AL. An approach to pose-based action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR; 2013.
82. Li Q, Wu J, Tu Z. Harvesting mid-level visual concepts from large-scale Internet images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR; 2013.
83. Myeong H, Lee KM. Tensor-Based High-Order Semantic Relation Transfer For Semantic Scene Segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR; 2013.
84. Zhao G, Yuan J, Hua G. 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*, Portland, OR; 2013.
85. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: *Proceedings of the International Conference on Very Large Data Bases*, Santiago de Chile, Chile; 1994, 487–499.
86. Pei J, Han J, Lakshmanan LV. Mining frequent itemsets with convertible constraints. In: *Proceedings of the IEEE International Conference on Data Engineering*, Heidelberg, Germany; 2001, 433–442.
87. Yuan J, Wu Y, Yang M. From frequent itemsets to semantically meaningful visual patterns. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Jose, CA; 2007, 864–873.
88. Yuan J, Wu Y, Yang M. Discovery of collocation patterns: from visual words to visual phrases. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN; 2007, 1–8.

89. Zhao G, Yuan J. Mining and cropping common objects from images. In: *Proceedings of the ACM International Conference on Multimedia*, Firenze, Italy; 2010, 975–978.
90. Liu J, Liu Y. GRASP Recurring Patterns from a Single View. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR; 2013.
91. Wang X, Grimson E. Spatial latent Dirichlet allocation. In: *Advances in Neural Information Processing Systems*, Vancouver, BC; 2008.
92. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn* 2001, 42:177–196.
93. Gaussier E, Goutte C. Relation between PLSA and NMF and implications. In: *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, Salvador, Brazil; 2005, 601–602.
94. Ding C, Li T, Peng W. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput Stat Data Anal* 2008, 52:3913–3927.
95. Jiang F, Yuan J, Tsaftaris SA, Katsaggelos AK. Anomalous video event detection using spatiotemporal context. *Comput Vis Image Understand* 2011, 115: 323–333.
96. Cong Y, Yuan J, Liu J. Sparse reconstruction cost for abnormal event detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO; 2011, 3449–3456.
97. Dong G, Li J. Efficient mining of emerging patterns: discovering trends and differences. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California; 1999, 43–52.
98. Wiskott L, Sejnowski TJ. Slow feature analysis: unsupervised learning of invariances. *Neural Comput* 2002, 14:715–770.

