Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Review:

A review of object representation based on local features*

Jian CAO†1, Dian-hui MAO1, Qiang CAI1, Hai-sheng LI1, Jun-ping DU2

(1College of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

(2School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

†E-mail: caojian@th.btbu.edu.cn; caojian9527@sina.com

Received Dec. 29, 2012; Revision accepted May 8, 2013; Crosschecked June 6, 2013

Abstract: Object representation based on local features is a topical subject in the domain of image understanding and computer vision. We discuss the defects of global features in present methods and the advantages of local features in object recognition, and briefly explore state-of-the-art recognition methods using local features, especially the main approaches of local feature extraction and object representation. To clearly explain these methods, the problem of local feature extraction is divided into feature region detection, feature region description, and feature space optimization. The main components and merits of these steps are presented. Technologies for object presentation are classified into three types: vector space, sliding window, and structure relationship models. Future development trends are discussed briefly.

Key words: Object presentation, Local feature, Image understanding, Object recognition, Visual words
doi:10.1631/jzus.CIDE1303 **Document code:** A **CLC number:** TP391.41

1 Introduction

How to recognize objects in images automatically is one of the most exciting and difficult problems in image processing and pattern recognition. The performance of an object recognition system depends crucially on two issues: the representation of the objects, and the classification algorithm adopted. Object representation, as a key technology, has a profound influence on the selection of the classifier and the eventual performance.

In recent years, many studies have attempted to describe objects based on global features. But in most cases, obtaining global features of an object in images is difficult on account of noise. Noise not only degrades parts of an image, but also makes it difficult to cover the entire scene of a capturing media. In practical applications, image targets are commonly articulated or occluded.

Many studies have shown that the human visual system can decompose one object into many meaningful small pieces, and identify the classification through this local information. Therefore, computer vision tasks using local features have attracted much attention from the research community. Local features are distinctive and insensitive to occlusion, and it is unnecessary to segment the objects from the background. Recent work has concentrated on making these new local features invariant to image transformation and on using them to represent the objects.

There are four requirements when we use local features in these applications. First, the regions around the point of interest must be localized in position and scale. In general, in a scale-space search, points of interest will be placed at local peaks and then filtered. The preserved points are likely to remain stable under transformation. Second, a description of these regions should be built; in an ideal condition, the description should be distinctive

* Project supported by the National Basic Research Program (973) of China (No. 2012CB821206), the National Natural Science Foundation of China (No. 71201004), the Scientific Research Common Program of Beijing Municipal Commission of Education (No. KM201310011009), and the Funding Project for Innovation on Science, Technology and Graduate Education in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality (Nos. PXM2012_014213_000037 and PXM2012_014213_000079)

(reasonably differentiating a region around a point of interest), brief, and robust under transformations caused by changes in camera pose and lighting. Third, for eliminating irrelevant and redundant local features, the feature space should be optimized to make the design of the classifier simple and to improve prediction performance and computational efficiency. Finally, a representation model should be designed which can describe objects based on suitable components and distinguish the objects clearly from others.

2 Region detectors

In general, the procedures for local feature extraction include region detection and feature description. Many different methods have been proposed for detecting and describing local image regions. Compared with the design of classifiers, these techniques are more dependent on specific problems and the knowledge of the relevant area. Some features should be invariant when the scale, noise, or 3D viewpoint is changed.

To compute the descriptors, the detectors present the regions, whose sizes rely on the detection scales. Detectors use different image measurements and all are based on the traversal of image pixels. At present, the commonest region detection methods can be classified into three categories: dense sampling, sparse sampling, and other methods.

2.1 Dense sampling methods

The researchers using these methods generally consider that all the regions have some role in low-level processing for pattern recognition. With any loss of detail, the eventual performance may be greatly affected.

Ohba and Ikeuchi (1997) described a method for dividing the images into dense non-overlapping ‘Eigen Windows’, with each window regarded as a local feature region. Jurie and Triggs (2005) treated each pixel of an image as the center of one region, and based the representation on densely sampled patches. Dalal and Triggs (2005) and Zhu *et al.* (2006) obtained many feature regions on all pixels and different scales of the detection window, and showed that dense sampling followed by explicit discriminative feature selection achieved the desired effect. Zhang LB *et al.* (2012) classified images using visual words based on dense sampling methods.

Dense sampling methods are always used for sliding window models. The advantage is that there is no loss of detail, and very rich local information can be obtained. But many regions have little useful information, and may even play a role in interference. Feature space optimization should be applied if dense sampling methods are to be used in object recognition systems.

2.2 Sparse sampling methods

In these methods, point detectors provide regions for computing the descriptors. If not stated otherwise, the detection scale determines the size of the regions. The detection operators mentioned in the current literature can be classified into two kinds: shape-based and appearance-based detectors.

Shape-based detectors conform to the shape information of the images (such as borders, straight lines, or arcs) to ascertain the location of points of interest and feature regions. They are used mainly to determine targets which are easy to distinguish from the shape, such as rigid and non-jointed objects. van Gool *et al.* (1996) and Fergus *et al.* (2005) detected points of interest by analyzing the image edge information. They discussed approaches for implementing first- and second-order digital derivatives for the detection of edges in an image, such as the Sobel, Prewitt, Laplacian, Laplacian of a Gaussian (LoG), and Canny operators. For each image, Berg *et al.* (2005) extracted sparse oriented edge maps and computed features at locations which have high edge energy. Intuitively, corners are a kind of useful point of interest and the features based on them can provide good results in image understanding and target recognition (Cao *et al.*, 2011b; 2011c).

Appearance-based detectors search for some kind of stable and invariant points or regions in the grayscale mode of images. Harris-Laplace regions (Mikolajczyk and Schmid, 2001) are corner-like structures, in which points are detected based on a scale-adapted Harris function and selected in scale-space using an LoG operator. Hessian-Laplace regions (Mikolajczyk *et al.*, 2005a) are determined in space at the local maxima of the Hessian determinant and in scale in the same way as Harris-Laplace detectors. Harris- and Hessian-Affine regions (Mikolajczyk and Schmid, 2004) show fine effects in consideration of affine image transformation. Difference-of-Gaussians (DoG) regions (Lowe, 2004) are robust to rotation and scale changes, and a detector based on such regions is suitable for finding blob-like structures. The points are local scale-space maxima of the DoG. By adopting entropy over local intensity histograms, a salient region (SalReg) detector identifies local image regions that are non-predictable across scales (Kadir *et al.*, 2004). The maximally stable extremal regions (MSER) technique applies a watershed-like segmentation algorithm to a target image and can be seen as components of connected pixels. The average *x* and *y* pixel locations are computed to obtain the position of the regions (Matas *et al.*, 2004).

The number of regions obtained by sparse sampling methods is always between 200 and 300. There are far fewer points of interest than in the pixels of the original images. So, the next process can be greatly accelerated and the workload of feature optimization can be reduced appropriately. But most region detectors are related to the characteristics of the image. When applied to a common object recognition system, there may be some limitations in this step.

2.3 Other methods

In comparison with the detectors described above, Nowak *et al.* (2006) found that random sampling gives equal or better classifiers. Although complicated multi-scale detectors work well for small numbers of samples, the single most important factor affecting performance is the number of patches sampled from the test image. The detectors above cannot provide enough patches for the next step in image processing. Moosmann *et al.* (2008) described how to select the image block dynamically and the methods of extremely randomized clustering forests (ERC-Forests). Experiments showed that their performance is satisfactory.

The dense sampling, sparse sampling, and other methods are all based on scanning, analyzing, and computing whole input images. The differences are that dense sampling methods can obtain very rich local information and aggravate the burden of feature space optimization. Random sampling methods require a large number of samples in the training set. Sparse sampling methods, such as Harris-Laplace and DoG detectors, are currently widely used in a variety of image understanding and object classification systems. However, the effect of each sparse sampling method often depends greatly on the characteristics of the objects and the background of the images.

3 Local feature descriptors

After the regions have been detected by methods such as those discussed above, the resulting aggregate of pixels usually is represented and described in a form suitable for further computer processing. In general, representing a region involves two choices: (1) the region can be represented in terms of its external characteristics (its boundary), or (2) it can be represented in terms of its internal characteristics (the pixels comprising the region). For scale and affine invariance, all the regions are mapped to a circular region of constant radius. Obviously, using the external characteristics we could not distinguish one normalized region from another. Thus, the important task is to describe the region based on its internal characteristics.

Feature vectors are always used in practice to describe local information of the regions. They are designed to capture perceptually meaningful distinctions and provide good resistance to extraneous details, such as changes in illumination. From the perspective of techniques and applications, local feature descriptors can be classified into three kinds: distribution-based, spatial-frequency, and other descriptors.

3.1 Distribution-based descriptors

These descriptors represent different characteristics of appearance or shape by means of histograms. One of the simplest approaches is to use statistical moments of the region-based gray-level histogram to describe texture. An intensity domain spin image (SI) is a 2D histogram that encodes the distribution of brightness values in an affine-normalized patch. Lazebnik *et al.* (2003) introduced a texture representation which is suitable for recognizing objects of textured surfaces using SI. As an illumination invariant descriptor, the local binary pattern (LBP) texture has been proven according to the analysis of a physical model of moving shadows, and the character of LBP texture can be utilized for shadow detection in different scenes (Cao *et al.*, 2011a; Pan *et al.*, 2011).

Lowe (2004) proposed a scale invariant feature transform (SIFT) descriptor which is a 3D histogram of gradient locations and orientations. The resulting descriptor is insensitive to small geometric distortions and small errors when detecting the region. Based on the fact that computation in the construction of a SIFT descriptor is very complex, Zeng and Gu (2012) introduced a strategy in which the dominant orientation of pixel gray gradients is computed in a circular region, and then the circular region is partitioned into eight identical sector areas starting from the dominant orientation. This method can increase the average matching speed significantly, even if there is affine distortion, defocusing, rotation, scaling, or illumination variation in the images.

Gradient location-orientation histogram (GLOH) (Mikolajczyk and Schmid, 2005) and PCA-SIFT (Ke and Sukthankar, 2004) descriptors are both extensions of the SIFT descriptor. Principal component analysis (PCA) is used to reduce the dimensions and the vectors of an image gradient, which are computed within the support regions as in SIFT. Geometric histogram and shape context implement the same idea, and compute a histogram describing the edge distribution in a region. So, both methods have been successfully used in shape analysis of drawings because edges are reliable features in such images. A shape-token is constructed by pairing a reference shape primitive to its neighboring primitive, and the shape primitives are composed of line segments and ellipses. Chia *et al.* (2012) proposed the contour features of shape-tokens, and discussed how to describe and match them at a single scale or across multiple scales. The SURF descriptor (Bay *et al.*, 2008) is based on similar properties and obtained from a square region snap to the selected orientation. In a state-of-the-art approach, a vocabulary of 3D visual words was produced by quantizing 3D SURF local descriptors (Redondo-Cabrera *et al.*, 2012), which were computed on partial 3D shapes extracted from the point clouds.

3.2 Spatial-frequency descriptors

Many descriptors are based on the frequency content of an image. The image content can be decomposed into the basis functions by Fourier transform. It is difficult to adapt to a local approach, because the spatial relations between points are not explicit and the basis functions are infinite. The Gabor transform can solve these problems. However, many Gabor filters have been used to capture small changes in frequency and orientation. Choi *et al.* (2012) proposed color local Gabor wavelets (CLGWs) which are able to exploit the discriminative information derived from spatiochromatic texture patterns of different spectral channels. This color local texture feature has been used for face recognition. Gabor filters and wavelets are often used in the fields of texture analysis and categorization.

From a mathematical viewpoint, images are 2D arrays of intensity values with locally varying statistics that result from different combinations of abrupt features, like edges and contrasting homogeneous regions. Wavelet transforms are based on small waves which are of varying frequency and limited duration. This allows them to provide the equivalent of a musical score for an image, unlike the Fourier transform. Papageorgiou and Poggio (2000), Mohan *et al.* (2001), and Viola and Jones (2004) described image regions using Harr wavelets in the frequency domain, and presented algorithms for object detection combined with kernel methods.

3.3 Other descriptors

A series of image derivatives computed up to a given order approximates a point neighborhood. Complex filters are the differential descriptors derived from the family $K(x, y, \theta)=f(x, y)\mathrm{e}^{\mathrm{i}\theta}$, where θ is the orientation. For function $f(x, y)$, Baumberg (2000) used Gaussian derivatives, and Schaffalitzky and Zisserman (2002) applied a polynomial. Generalized moment invariants have been introduced by van Gool *et al.* (1996), for describing the multispectral nature of image data. The invariants combine central moments defined by $M_{pq}^{a}=\iint_{\Omega} x^{p} y^{q}\left(I(x, y)\right)^{a} \mathrm{d}x\mathrm{d}y$ of order $p+q$ and degree a. The moments characterize shape and intensity distribution in a region Ω. They are independent and can be easily computed for any order and degree. However, the moments of high order and degree are sensitive to small geometric and photometric distortions. Computing the invariants reduces the number of dimensions.

A new descriptor, called illumination invariant multiscale autoconvolution (MSA) moments, was proposed by Ding *et al.* (2012). It performs very well under strong non-affine transformation in 3D object recognition tasks. Chen and Gleason (2012) developed a new set of invariant moments based on the ridgelet function, which is good at capturing line features in a pattern image. These descriptors are therefore more suitable for color images where the invariants can be computed for each color channel and between the channels.

With advances in technology, a large number of descriptors for local features have been proposed. Each descriptor can perform well in a certain suitable area, and their characteristics are not closely related to the detection methods. Overall, SIFT and GLOH descriptors are among the best for accuracy and consistency. SIFT is significantly faster than GLOH and others, which is important for real time machine vision applications. The shape context descriptor has been successfully used in shape analysis of drawings in which edges are reliable features, but rarely used in the classification of texture images and non-rigid objects. Generalized moment invariants and complex filters perform better than other low-dimensional descriptors in practical applications.

4 Feature space optimization

Intuitively, more local features mean more discriminative power in object recognition. However, this is not always true in practice. Many features might be irrelevant and possibly detrimental to classification. Meanwhile, there is much redundancy among features (Yu and Liu, 2004). Irrelevant and redundant features always slow down the learning algorithm and cause it to overfit the training data. The elimination of such features, could not only simplify the design of the classifier, but also improve its prediction performance and computational efficiency.

Many researchers start facing the problems of learning algorithms with high-dimensional data sets. The algorithms can be divided into two categories: feature selection and feature transform.

4.1 Feature selection

Feature selection methods aim to select a subset of relevant features to achieve similar or even better results in object recognition than using the original feature set. Existing feature selection algorithms can be divided into four categories: filter, wrapper, embedded, and mixture approaches.

Unlike filter approaches, which are arguably less expensive to compute and more general, wrapper approaches use a learning algorithm in the evaluation procedure and achieve better results. Relief is a popular selection method that filters out irrelevant features based on the nearest neighbor approach. Recursive feature elimination support vector machine (RFE-SVM) is a widespread method which selects useful features while training the SVM classifier (Guyon *et al.*, 2002). The genetic programming relevance measure (GPRM) has been used by Neshatian and Zhang (2009) for evaluating and ranking subsets of features in binary classification tasks. GPRM is efficient in terms of feature selection.

Feature selection based on some elements of information theory (Agarwal *et al.*, 2004; Jurie and Triggs, 2005; Mikolajczyk *et al.*, 2005b) is a hot spot in recent research, such as image frequency, term strength, χ^2 statistics, information gain (IG), and mutual information (MI). Liu *et al.* (2009) proposed a new feature selection algorithm based on dynamic mutual information, which is estimated only on unlabeled instances. A correlation-based method was proposed by Yu and Liu (2004) for relevance and redundancy analysis. By comparison with other representative methods, it conducts an empirical study of its efficiency and effectiveness. Javed *et al.* (2012) developed a new feature selection algorithm, called class-dependent density-based feature elimination (CDFE), which uses a classifier for the selection of the final subset instead of a threshold value provided by the users.

4.2 Feature transform

Feature transform methods are intended to produce a new set of features from the original features by means of combination or transformation technologies. PCA, independent component analysis (ICA), and linear discriminant analysis (LDA) are three renowned linear algorithms, which have been widely used, because they are simple and effective.

PCA (Shlens, 2009) picks the directions that maximize feature variance, and therefore describes the ‘most important’ changes in the data. The goal of ICA (Hyvarinen and Oja, 2000) is to find a linear representation of non-Gaussian data to ensure that the components are statistically independent, or as independent as possible. This kind of representation appears to capture the essential structure of the data in local feature extraction. Unlike PCA, LDA (Dai and Yuen, 2003; Yang and Yang, 2003) uses class labels for searching the directions along which to separate the class means relative to the sum of the class variances. LDA maximizes the ratio of between-class scatter to within-class scatter. Visually, lower dimensional descriptions of the data obtained can push the class members together, and pull the members of different classes apart.

Nonlinear feature transform algorithms include Isomap (Geng *et al.*, 2005) and locally linear embedding (LLE) (Saul and Roweis, 2003). Almost all of the linear methods can be extended to the kernel space, based on which the kernel Fisher discriminant analysis (FDA), kernel locality preserving projections (LPP), and kernel direct LDA algorithms were proposed. 2D-PCA (Yang *et al.*, 2004) and 2D-LDA (Yang *et al.*, 2005) use 2D image matrices rather than 1D vectors. In feature extraction, we need not transform the image matrix into a vector in advance. Instead, we use the original image matrices to directly construct an image covariance matrix.

Feature spaces are usually designed to capture perceptually meaningful distinctions while providing good resistance to extraneous details, such as features with interferential information. In a properly chosen representation method, the distances in the feature space appear to coincide approximately with distinguishability concerning perceptual 'noise'. Because coding dense regions more finely is wasteful, a perceptually efficient coding should have approximately uniform cell size, or at least actualize a lower bound on the cell size.

All feature space optimization algorithms aim to obtain a small number of features to give similar or even better recognition results than the original features. Feature selection methods are often used in conjunction with feature transform (Li *et al.*, 2012). For example, we can first pick a subset of the original features, and then transform this subset in the next steps. In recent years, many optimization algorithms (Hancock and Mamitsuka, 2012; He *et al.*, 2012; Zhang LJ *et al.*, 2012) for feature space have been proposed, and most of them have been used in object detection or classification.

5 Representation model

Three common pattern arrangements are used in practice: vectors (for quantitative descriptions), strings, and trees (for structural descriptions). Once the objects or images have been described based on local features using these pattern arrangements, the complex problems of generic visual recognition can be reduced to multi-class supervised learning in pattern recognition fields. Then classifiers such as Naïve Bayes, neural network, and support vector machine begin to carry out two separate steps, training and testing, to ensure the classification of unlabeled objects.

An ideal object representation method should be able to choose suitable components to describe the contents of one object. It should also have the ability to distinguish easily one object from another. Object representation methods based on local features can be classified into three kinds: vector space models (VSM), sliding window models (SWM), and structure relationship models (SRM).

5.1 Vector space models

VSM was proposed by Salton *et al.* (1975) and has been widely used for text categorization (Cristianini *et al.*, 2002; Tong and Koller, 2002). Recently, researchers wanted to adapt this text categorization approach to visual recognition, and renamed it bag-of-words (Nakayama *et al.*, 2010), codebook (Jurie and Triggs, 2005), or visual vocabulary (Perronnin and Dance, 2007).

The vocabulary should be large enough to distinguish relevant changes in image parts, but not so large as to distinguish irrelevant variations such as noise. In visual recognition, 'words' does not necessarily have a repeatable meaning such as 'noses' or 'wings', nor is there an evident best choice of bag-of-words. There are two ways of extracting feature vectors of the images from the visual vocabulary: histograms and binary indicator vectors. In histograms, the feature vector is simply the normalized histogram (scaled to total sum 1) of occurrence counts for the different visual words. However, for indicator vectors, each component is 1 if one region with the visual word occurs in the target image, and is 0 otherwise.

Mikolajczyk *et al.* (2006) used a generative model to recognize and locate multiple object classes simultaneously. A codebook representation was presented in this method in which appearance clusters based on edge features were shared among several target classes. Wang *et al.* (2012) proposed a visual word soft-histogram for image object representation based on statistical modeling and discriminative learning of visual words. Fernando *et al.* (2012) proposed a new logistic regression-based fusion algorithm, called LRFF, which capitalizes on different cues without being tied to any of them. Meanwhile, they projected a new marginalized kernel by means of the output from the regression model.

Despite the fact that VSM captures only local statistics and ignores geometric relationships, methods based on VSM perform surprisingly well. They are elastic and quite easy to build. They can capture a great proportion of the complex statistics of the images and target classes in a convenient local mode, and they are robust to partial occlusions, geometric deformations, and illumination changes.

5.2 Sliding window models

In SWM, image objects are represented using components based on visual words, together with spatial relation scanning among the components (Agarwal *et al.*, 2004). If we have trained a classifier to have the ability to distinguish the images between positive and negative, the target objects can be detected by moving a sliding window over the image and classifying each window as positive or negative. However, because of the invariance of the classifier to slight changes in a target, several windows in the vicinity of the target in the image will be classified as positive. This will induce multiple detections for only one target.

Dalal and Triggs (2005) and Zhu *et al.* (2006) applied SWM to detect pedestrians in images, and the detection results demonstrated that this strategy can achieve good results in practical applications. Zhang LB *et al.* (2012) investigated the potential strength of the relative position information of visual words and proposed a new kind of representation named bag-of-phrases. Bilen *et al.* (2012) considered localization, in the form of a window, as a latent variable that is learnt jointly with other classification model parameters. This framework generalizes feature localization using multiple cues, which is beneficial for classification. Chen *et al.* (2012) proposed a generalized hierarchical matching (GHM) framework based on an SWM for object classification. The methods based on this flexible and general scheme can take advantage of any useful side information in the pattern recognition framework.

5.3 Structure relationship models

Some technologies match the targets based on a cost value for the deformation needed to transform a prototypical model to coincide with the object. Active appearance models (Belongie *et al.*, 2002) use two steps: first, deform the target to a mean shape, and second, estimate the combined modes of variation of the concatenated shape and texture models. If this algorithm is initialized with a close estimate of the target's location and size, a remarkably good performance can be obtained within a few iterations, even for deformable objects.

Weber *et al.* (2000) obtained the assembly of object parts by modeling their joint spatial probability distribution. This scheme has been extended to scale-invariant object parts and the modified result is called the constellation model (Fergus *et al.*, 2005). This has been successfully demonstrated with several image categories. Initially, it modeled the relative part locations based on a fully connected graph. In later versions, a simpler star topology was used, instead of the fully-connected graph, which can cope with many more parts using efficient inference algorithms.

The model of Agarwal *et al.* (2004) learns the spatial configurations between pairs of object parts on the basis of a classifier. Nevertheless, the learning algorithm needs a large number of training examples and the repeated observation of occurrences between the same parts in similar spatial relations. Leibe *et al.* (2008) presented the implicit shape model (ISM). This model combines a lot of automatically selected parts detected by a point of interest operator flexibly based on a star topology.

These three models are used for object representation and are suitable for different applications. VSM aims to obtain a histogram on the basis of the occurrences of particular target patterns in a given image. This model has advantages such as simplicity, computational efficiency, and invariance to affine transformations, even when local occlusion, lighting, and intra-class variations occur. SWM performs well on a training set which is large in scale and with densely sampled regions. Object recognition based on this model is always in conjunction with multi-resolution analysis and adaboost technologies. SRM captures structure relationships of the object parts, and is useful for its expansiveness. As many technologies of pattern recognition can be used in SRM, this model is able to play an even greater role in object detection and segmentation.

6 Conclusions

Local features computed on regions of interest have proven to be very successful in real-world applications such as wide baseline matching, image retrieval, building panoramas, machine vision, video data mining, and texture classification. They are characteristic, effective to occlusion, relatively insensitive to changes in viewpoint, and not based on segmentation of the whole target from the background. Recently, remarkable progress has been made in object representation on the basis of local features. Meanwhile, several robust algorithms have been developed for recognizing targets in real time in simple scenes. Nevertheless, it can be seen from the literature reviewed in this survey that many assumptions have been used to simplify the recognition problem, such as high contrast between the background and foreground, few local occlusions, and constant illumination. These conditions are scarcely possible in many real scenes and render the methods useless for many applications. Object recognition and associated problems of feature extraction, feature space optimization, and object representation are active fields of research and new solutions are continuously being put forward.

One challenge in object representation is how to integrate multiple cues and combine several local features for multi-category discrimination (Chen *et al.*, 2012; Chia *et al.*, 2012). Another direction for future research is to describe the object based on global features in conjunction with local features (Bilen *et al.*, 2012). It can be advantageous to work on several rescaled versions of the image, owing to high computational efficiency. Finally, in a number of practical applications, the image targets should be described and recognized from multiple viewpoints. We can obtain 2D views of a 3D model in different view regions, and use these projections to simulate the images of the target with various poses or changes in 3D viewpoint (Cao *et al.*, 2011b). Overall, we suppose that whenever possible, full use should be made of additional information, especially prior and contextual information, to coordinate the descriptor and the classifier with the particular scene used.

In this paper, we have presented an extensive survey of object representation methods based on local features and given a brief review of related topics. We believe that this survey of object representation based on local features with an extensive literature review, can give valuable insights into this important research topic and encourage new research.

References

Agarwal, S., Awan, A., Roth, D., 2004. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**(11):1475-1490. [doi:10.1109/TPAMI.2004.108]

Baumberg, A., 2000. Reliable Feature Matching across Widely Separated Views. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.774-781. [doi:10.1109/CVPR.2000.855899]

Bay, H., Ess, A., Tuytelaars, T., Gool, L.V., 2008. SURF: speeded up robust features. *Comput. Vis. Image Understand.*, **110**(3):346-359. [doi:10.1016/j.cviu.2007.09.014]

Belongie, S., Malik, J., Puchiza, J., 2002. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**(4):509-522. [doi:10.1109/34.993558]

Berg, A.C., Berg, T.L., Malik, J., 2005. Shape Matching and Object Recognition Using Low Distortion Correspondences. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.26-33. [doi:10.1109/CVPR.2005.320]

Bilen, H., Namboodiri, V.P., van Gool, L.J., 2012. Classification with global, local and shared features. *LNCS*, **7476**: 134-143. [doi:10.1007/978-3-642-32717-9_14]

Cao, J., Chen, H.Q., Zhang, K., Niu, C.F., 2011a. Moving cast shadow detection based on region color and texture. *Robot*, **33**(5):628-633 (in Chinese).

Cao, J., Liu, Q.X., Gao, C.X., Liu, Y.S., 2011b. Object recognition with corner-based feature. *Trans. Beijing Inst. Technol.*, **31**(3):308-312 (in Chinese).

Cao, J., Chen, H.Q., Mao, M.Y., 2011c. Optimization Algorithms for Local Features. Int. Conf. on Automation, Communication, Architectonics and Materials, p.921-924. [doi:10.4028/www.scientific.net/AMR.225-226.921]

Chen, G.Y., Gleason, S., 2012. Ridgelet Moment Invariants for Pattern Recognition. Proc. 25th IEEE Canadian Conf. on Electrical & Computer Engineering, p.1-4. [doi:10.1109/CCECE.2012.6335061]

Chen, Q., Song, Z., Hua, Y., Huang, Z.Y., Yan, S.C., 2012. Hierarchical Matching with Side Information for Image Classification. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.3426-3433. [doi:10.1109/CVPR.2012.6248083]

Chia, A.Y., Rajan, D., Leung, M.K., Rahardja, S., 2012. Object recognition by discriminative combinations of line segments, ellipses, and appearance features. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(9):1758-1772. [doi:10.1109/TPAMI.2011.220]

Choi, J.Y., Ro, Y.M., Plataniotis, K.N., 2012. Color local texture features for color face recognition. *IEEE Trans. Image Process.*, **21**(3):1366-1380. [doi:10.1109/TIP.2011.2168413]

Cristianini, N., Shawe-Taylor, J., Lodhi, H., 2002. Latent semantic kernels. *J. Intell. Inf. Syst.*, **18**(2/3):127-152. [doi:10.1023/A:1013625426931]

Dai, D.Q., Yuen, P.C., 2003. Regularized discriminant analysis and its application to face recognition. *Pattern Recogn.*, **36**(3):845-847. [doi:10.1016/S0031-3203(02)00092-4]

Dalal, N., Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.886-893. [doi:10.1109/CVPR.2005.177]

Ding, H., Li, X.D., Zhao, H.J., Xiao, W., 2012. A New Generalized Affine Moment Invariants for Shape Retrieval and Object Recognition. Proc. IEEE Int. Symp. on Instrumentation and Control Technology, p.137-142. [doi:10.1109/ISICT.2012.6291609]

Fergus, R., Perona, P., Zisserman, A., 2005. A Sparse Object Category Model for Efficient Learning and Exhaustive Recognition. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.380-387. [doi:10.1109/CVPR.2005.47]

Fernando, B., Fromont, E., Muselet, D., Sebban, M., 2012. Discriminative Feature Fusion for Image Classification. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.3434-3441. [doi:10.1109/CVPR.2012.6248084]

Geng, X., Zhan, D.C., Zhou, Z.H., 2005. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Trans. Syst. Man Cybern. B*, **35**(6):1098-1107. [doi:10.1109/TSMCB.2005.850151]

Guyon, I., Watson, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**(1-3):389-422. [doi:10.1023/A:1012487302797]

Hancock, T., Mamitsuka, H., 2012. Boosted network classifiers for local feature selection. *IEEE Trans. Neur. Networks Learn. Syst.*, **23**(11):1767-1778. [doi:10.1109/TNNLS.2012.2214057]

He, R., Tan, T.N., Wang, L., Zheng, W.S., 2012. Regularized Correntropy for Robust Feature Selection. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.2504-2511. [doi:10.1109/CVPR.2012.6247966]

Hyvarinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. *Neur. Networks*, **13**(4-5):411-430. [doi:10.1016/S0893-6080(00)00026-5]

Javed, K., Babri, H.A., Saeed, M., 2012. Feature selection based on class-dependent densities for high-dimensional binary data. *IEEE Trans. Knowl. Data Eng.*, **24**(3):465-477. [doi:10.1109/TKDE.2010.263]

Jurie, F., Triggs, B., 2005. Creating Efficient Codebooks for Visual Recognition. Proc. 10th IEEE Int. Conf. on Computer Vision, p.604-610. [doi:10.1109/ICCV.2005.66]

Kadir, T., Zisserman, A., Brady, M., 2004. An affine invariant salient region detector. *LNCS*, **3021**:228-241. [doi:10.1007/978-3-540-24670-1_18]

Ke, Y., Sukthankar, R., 2004. PCA-SIFT: a More Distinctive Representation for Local Image Descriptors. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.511-517.

Lazebnik, S., Schmid, C., Ponce, J., 2003. A Sparse Texture Representation Using Affine-Invariant Regions. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.319-324. [doi:10.1109/CVPR.2003.1211486]

Leibe, B., Leonardis, A., Schiele, B., 2008. Robust object detection with interleaved categorization and segmentation. *Int. J. Comput. Vis.*, **77**(1-3):259-289. [doi:10.1007/s11263-007-0095-3]

Li, C.S., Liu, Q.S., Liu, J., Lu, H.Q., 2012. Learning Ordinal Discriminative Features for Age Estimation. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.2570-2577. [doi:10.1109/CVPR.2012.6247975]

Liu, H.W., Sun, J.G., Liu, L., Zhang, H.J., 2009. Feature selection with dynamic mutual information. *Pattern Recogn.*, **42**(7):1330-1339. [doi:10.1016/j.patcog.2008.10.028]

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, **60**(2):91-110. [doi:10.1023/B:VISI.0000029664.99615.94]

Matas, J., Chum, O., Urban, M., Pajdla, T., 2004. Robust wide baseline stereo from maximally stable extremal regions. *Image Vis. Comput.*, **22**(10):761-767. [doi:10.1016/j.imavis.2004.02.006]

Mikolajczyk, K., Schmid, C., 2001. Indexing Based on Scale Invariant Interest Points. Proc. 8th Int. Conf. on Computer Vision, p.525-531. [doi:10.1109/ICCV.2001.937561]

Mikolajczyk, K., Schmid, C., 2004. Scale & affine invariant interest point detectors. *Int. J. Comput. Vis.*, **60**(1):63-86. [doi:10.1023/B:VISI.0000027790.02288.f2]

Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**(10):1615-1630. [doi:10.1109/TPAMI.2005.188]

Mikolajczyk, K., Leibe, B., Schiele, B., 2005a. Local Features for Object Class Recognition. Proc. 10th IEEE Int. Conf. on Computer Vision, p.1792-1799. [doi:10.1109/ICCV.2005.146]

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., van Gool, L., 2005b. A comparison of affine region detectors. *Int. J. Comput. Vis.*, **65**(1-2):43-72. [doi:10.1007/s11263-005-3848-x]

Mikolajczyk, K., Leibe, B., Schiele, B., 2006. Multiple Object Class Detection with a Generative Model. Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, p.26-36. [doi:10.1109/CVPR.2006.202]

Mohan, A., Papageorgiou, C., Poggio, T., 2001. Example-based object detection in images by components. *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**(4):349-361. [doi:10.1109/34.917571]

Moosmann, F., Nowak, E., Jurie, F., 2008. Randomized clustering forests for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, **30**(9):1632-1646. [doi:10.1109/TPAMI.2007.70822]

Nakayama, H., Harada, T., Kuniyoshi, Y., 2010. Global Gaussian Approach for Scene Categorization Using Information Geometry. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.2336-2343. [doi:10.1109/CVPR.2010.5539921]

Neshatian, K., Zhang, M., 2009. Genetic programming for feature subset ranking in binary classification problems. *LNCS*, **5481**:121-132. [doi:10.1007/978-3-642-01181-8_11]

Nowak, E., Jurie, F., Triggs, B., 2006. Sampling strategies for bag-of-features image classification. *LNCS*, **3954**:490-503. [doi:10.1007/11744085_38]

Ohba, K., Ikeuchi, K., 1997. Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *IEEE Trans. Pattern Anal. Mach. Intell.*, **19**(9):1043-1047. [doi:10.1109/34.615453]

Pan, H., Li, X.B., Jin, L.Z., Xia, L.Z., 2011. Object description and recognition using multiscale geometric analysis. *J. Infrar. Millim. Waves*, **30**(1):85-90.

Papageorgiou, C., Poggio, T., 2000. A trainable system for object detection. *Int. J. Comput. Vis.*, **38**(1):15-33. [doi:10.1023/A:1008162616689]

Perronnin, F., Dance, C., 2007. Fisher Kernels on Visual Vocabularies for Image Categorization. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8. [doi:10.1109/CVPR.2007.383266]

Redondo-Cabrera, C., Lopez-Sastre, R.J., Acevedo-Rodriguez, J., Maldonado-Bascon, S., 2012. SURFing the Point Clouds: Selective 3D Spatial Pyramids for Category-Level Object Recognition. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.3458-3465. [doi:10.1109/CVPR.2012.6248087]

Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. *Commun. ACM*, **18**(11):613-620. [doi:10.1145/361219.361220]

Saul, L.K., Roweis, S.T., 2003. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, **4**(2):119-155.

Schaffalitzky, F., Zisserman, A., 2002. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". *LNCS*, **2350**:414-431. [doi:10.1007/3-540-47969-4_28]

Shlens, J., 2009. A Tutorial on Principal Component Analysis. Center for Neural Science, New York University, New York City, NY. Available from http://www.snl.salk.edu/~shlens/pca.pdf

Tong, S., Koller, D., 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, **2**:45-66.

van Gool, L., Moons, T., Ungureanu, D., 1996. Affine/photometric invariants for planar intensity patterns. *LNCS*, **1064**:642-651. [doi:10.1007/BFb0015574]

Viola, P., Jones, M.J., 2004. Robust real-time face detection. *Int. J. Comput. Vis.*, **57**(2):137-154. [doi:10.1023/B:VISI.0000013087.49260.fb]

Wang, Y.J., Liu, X.B., Jia, Y.D., 2012. Visual word soft-histogram for image representation. *J. Softw.*, **23**(7):1787-1795 (in Chinese).

Weber, M., Welling, M., Perona, P., 2000. Unsupervised Learning of Models for Recognition. European Conf. on Computer Vision, p.18-32.

Yang, J., Yang, J.Y., 2003. Why can LDA be performed in PCA transformed space. *Pattern Recogn.*, **36**(2):563-566. [doi:10.1016/S0031-3203(02)00048-1]

Yang, J., Zhang, D., Frangi, A.F., Yang, J.Y., 2004. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**(1):131-137. [doi:10.1109/TPAMI.2004.1261097]

Yang, J., Zhang, D., Yong, X., Yang, J.Y., 2005. Two-dimensional discriminant transform for face recognition. *Pattern Recogn.*, **38**(7):1125-1129. [doi:10.1016/j.patcog.2004.11.019]

Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.*, **5**:1205-1224.

Zeng, L., Gu, D.L., 2012. A SIFT feature descriptor based on sector area partitioning. *Acta Autom. Sin.*, **38**(9):1513-1519 (in Chinese).

Zhang, L.B., Wang, C.H., Xiao, B.H., Shao, Y.X., 2012. Image representation using bag-of-phrases. *Acta Autom. Sin.*, **38**(1):46-54 (in Chinese).

Zhang, L.J., Chen, C., Bu, J.J., He, X.F., 2012. A unified feature and instance selection framework using optimum experimental design. *IEEE Trans. Image Process.*, **21**(5):2379-2388. [doi:10.1109/TIP.2012.2183879]

Zhu, Q., Avidan, S., Yeh, M.C., Cheng, K.T., 2006. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1491-1498. [doi:10.1109/CVPR.2006.119]