EXTRACTING INTERESTING POINTS FROM IMAGES

KORNEL BERTÓK, GERGŐ PÓLYI[1]

Faculty of Informatics, University of Debrecen, Debrecen, Hungary

[1]bertok.kornel@inf.unideb.hu, polyigergo@freemail.hu

1 Introduction

The purpose of this work is to create a mobile application which will allow a user to discover information about a given building or landmark. After determining in which building the user is interested, specific information can be returned about this building and/or the surrounding area. In particular, this technology can be used as part of a tourist application based in a city space, which will provide both historical and current information about the area.

So in this paper, we study the problem of building recognition. As an instance of the recognition problem, this domain is interesting since the class of buildings posses many similarities, while at the same time calls for techniques which are capable of fine discrimination between different instances of the class.

One of the central issues pertinent to the recognition problem is the choice of suitable representation of the class and its scalability to a large number of exemplars. There is a large amount of literature on general object recognition. The existing methods exploit geometric and/or appearance information, consider part based models or more holistic representations. In the context of the presented work we will review some related works which obtain the desired representations from image appearance, computing both global and local image features.

Global approaches typically consider the entire image as a point in the high-dimensional space and model the changes in the appearance as a function of viewpoint using subspace methods. Given the subspace representation the pose of the camera can be obtained by spline interpolation method, exploiting the continuity of the mapping between the object appearance and continuously changing viewpoint.

Alternative global representations proposed in the past include responses to banks of filters, multidimensional receptive field histograms and color histograms. These representations do not encode spatial information inherent in the image. Although quite robust to changes in viewpoint and/or scale, they are often not very discriminative. Partial means of encoding the spatial information can be obtained by computing the global descriptors over different image regions separately.

Alternative representations in terms of local features have become very effective in the context of different object/category recognition problems. In this case the descriptors are computed only over local image regions, whose location is first determined using various saliency measures. These representations perform favorably in the presence of large amount of clutter and changes in viewpoint. The representatives of local image descriptors include scale invariant features and their associated descriptors, which are robust with respect to affine transformations.

1.1 Outline

This study concentrate only for extracting interesting points from images, the feature matching and indexing is not the part of it. In this paper, we propose the base of a building recognition problem by a two stage hierarchical scheme. The first stage is comprised to extract simultaneously efficient local image descriptors to get a small number of best candidate images from the database. These candidates are chosen for the second recognition stage, which rectifies the images along the local image descriptors. Then global image descriptors will be extracted simultaneously from the normalized images to make the final decision.

2 Extracting Local Image Features

Most object recognition systems tend to use either global image features, which describe an image as a whole, or local features, which represent image patches. Local features are computed at multiple points in the image and are consequently more robust to occlusion and clutter. However, they may require specialized classification algorithms to handle cases in which there are a variable number of feature vectors per image.

Most local features represent texture in an image patch. For example, SIFT features use histograms of gradient orientations. One advantage of using local features is that they may be used to recognize the object despite significant clutter and occlusion. They also do not require a segmentation of the object from the background, unlike many texture features, or representations of the object's boundary (shape features).

2.1 SIFT: Scale Invariant Feature Transform

The SIFT [1,2] approach transforms an image into a large collection of local feature vectors. Each of these is invariant to image translation, scaling, and rotation, and partially invariant to illumination changes and affine or 3D projection. Previous approaches to local feature generation lacked invariance to scale and were more sensitive to projective distortion and illumination change. The SIFT features share a number of properties in common with the responses of neurons in inferior temporal (IT) cortex in primate vision.

The author also describes improved approaches to indexing and model verification. The scale-invariant features are efficiently identified by using a staged filtering approach. The first stage identifies key locations in scale space by looking for locations that are maxima or minima of a difference-of-Gaussian function. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame. The features achieve partial invariance to local variations, such as affine or 3D projections, by blurring image gradient locations. This approach is based on a model of the behavior of complex cells in the cerebral cortex of mammalian vision. The resulting feature vectors are called SIFT keys. In the current implementation, each image generates on the order of 1000 SIFT keys, a process that requires less than 1 second of computation time.

The SIFT keys derived from an image are used in a nearest-neighbor approach to indexing to identify candidate object models. Collections of keys that agree on a potential model pose are first identified through a Hough transform hash table, and then through a least-squares fit to a final estimate of model parameters. When at least 3 keys agree on the model parameters with low residual, there is strong evidence for the presence of the object. Since there may be dozens of SIFT keys in the image of a typical object, it is possible to have substantial levels of occlusion in the image and yet retain high levels of reliability.

The current object models are represented as 2D locations of SIFT keys that can undergo affine projection. Sufficient variation in feature location is allowed to recognize perspective projection of planar shapes at up to a 60 degree rotation away from the camera or to allow up to a 20 degree rotation of a 3D object.

2.2 SURF: Speeded Up Robust Feature

SURF is a robust local feature detector, first presented by Herbert Bay et al. in 2006 [3], that can be used in computer vision tasks like object recognition or 3D reconstruction. It is partly inspired by the SIFT descriptor. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT. SURF is based on sums of 2D Haar wavelet responses and makes an efficient use of integral images. It uses an integer approximation to the determinant of Hessian blob detector, which can be computed extremely quickly with an integral image (3 integer operations). For features, it uses the sum of the Haar wavelet response around the point of interest. Again, these can be computed with the aid of the integral image.

When working with local features, a first issue that needs to be settled is the required level of invariance. Clearly, this depends on the expected geometric and photometric deformations, which in turn are determined by the possible changes in viewing conditions. SURF focus on invariant of scale

and image rotation. These seem to offer a good compromise between feature complexity and robustness to commonly occurring deformations. Skew, anisotropic scaling and perspective effects are assumed to be second-order effects, that are covered to some degree by the overall robustness of the descriptor. As also claimed by Lowe [2], the additional complexity of full affine-invariant features often has a negative impact on their robustness and does not pay off, unless really large viewpoint changes are to be expected. In some cases, even rotation invariance can be left out, resulting in a scale-invariant only version of this descriptor, which is referred to as "upright SURF" (U-SURF). Indeed, in quite a few applications, like mobile robot navigation or visual tourist guiding, the camera often only rotates about the vertical axis. The benefit of avoiding the overkill of rotation invariance in such cases is not only increased speed, but also increased discriminative power.

2.3 FAST: Features from Accelerated Segment Test

FAST [4] is a corner detection method, which can be used to extract feature points and later used to track and map objects in many computer vision tasks. The advantage of FAST corner detector is its computational efficiency. FAST corner detector is very suitable for real-time video processing application because of high-speed performance. The FAST corner detector functions on a simple, but surprisingly effective algorithm. For any given pixel, a circle of pixels around that pixel is examined. If a large continuous chain of pixels in that circle are all significantly greater than or less than the current pixel, then that pixel is classified as a corner. The parameters to FAST are the required chain length and the amount by which the chain must be greater than or less than the query pixel.

It should not be surprising that clumps of corners are normally found, instead of one. Described in the paper [5] a metric can be used to determine corner strength and then that metric can be used as a non-maximal suppression step to find isolated corners. The FAST algorithm is thus a two-pass algorithm. In the first pass over the image, each pixel is examined and the corner strength metric is calculated if that pixel is a corner. A simple acceleration technique arises from the assumption that most pixels are not corners. If a large chain does in fact exist in the circle around the pixel, then $3/4$ of the cardinal directions must be in the chain ($2/4$ for chain lengths less than 12). Verifying this property eliminates a large number of pixels without the full circle computation. The second pass is the non-maximal suppression pass. This pass iterates over all corners from the first pass. If that corner is not a local maximum of a $N \times N$ neighborhood in terms of corner strength, it is discarded.

2.4 ORB: Oriented FAST and Rotated BRIEF

Feature matching is at the base of many computer vision problems, such as object recognition or structure from motion. Current methods rely on costly descriptors for detection and matching. ORB [6] was proposed as a very fast binary descriptor based on BRIEF, called ORB, which is rotation

invariant and resistant to noise. It was shown through experiments how ORB is at two orders of magnitude faster than SIFT, while performing as well in many situations.

The SIFT keypoint detector and descriptor [1], although over a decade old, have proven remarkably successful in a number of applications using visual features, including object recognition, image stitching, visual mapping, etc. However, it imposes a large computational burden, especially for real-time systems such as visual odometry, or for low-power devices such as cell phones. This has led to an intensive search for replacements with lower computation cost; arguably the best of these is SURF [3]. There has also been research aimed at speeding up the computation of SIFT, most notably with GPU devices.

ORB is a computationally-efficient replacement to SIFT that has similar matching performance, is less affected by image noise, and is capable of being used for real-time performance. ORB descriptor performs as well as SIFT on these tasks (and better than SURF), while being almost two orders of magnitude faster.

ORB builds on the well-known FAST keypoint detector [4,5] and the recently-developed BRIEF descriptor [7]; for this reason we call it ORB (Oriented FAST and Rotated BRIEF). Both these techniques are attractive because of their good performance and low cost. An additional benefit of ORB is that it is free from the licensing restrictions of SIFT and SURF.

2.5 MSER: Maximally Stable Extremal Regions

MSER [8] is used as a method of blob detection in images. This technique was proposed to find correspondences between image elements from two images with different viewpoints. This method of extracting a comprehensive number of corresponding image elements contributes to the wide-baseline matching, and it has led to better stereo matching and object recognition algorithms.

MSER is useful for the wide-baseline stereo problem, i.e. the problem of establishing correspondences between a pair of images taken from different viewpoints. A new set of image elements that are put into correspondence, the so called extremal regions, is introduced. Extremal regions possess highly desirable properties: the set is closed under 1. continuous (and thus projective) transformation of image coordinates and 2. monotonic transformation of image intensities. An efficient (near linear complexity) and practically fast detection algorithm (near frame rate) is realized for an affine-invariant stable subset of extremal regions, the maximally stable extremal regions (MSER).

MSER is a new robust similarity measure for establishing tentative correspondences. The robustness ensures that invariants from multiple measurement regions (regions obtained by invariant constructions from extremal regions), some that are significantly larger (and hence discriminative) than the MSERs, may be used to establish tentative correspondences. The high utility

of MSERs, multiple measurement regions and the robust metric is demonstrated in wide-baseline experiments on image pairs from both indoor and outdoor scenes. Significant change of scale (3.5×), illumination conditions, out-of-plane rotation, occlusion, locally anisotropic scale change and 3D translation of the viewpoint are all present in the test problems. Good estimates of epipolar geometry (average distance from corresponding points to the epipolar line below 0.09 of the inter-pixel distance) are obtained.

2.6 Star Feature Detector

Star Feature Detector is derived from CenSurE (Center Surrounded Extrema) detector [9]. While CenSurE uses polygons such as Square, Hexagon and Octagons as a more computable alternative to circle. Star simulates the circle with two overlapping squares: one is upright and the other one is 45° degree rotated. These polygons are bi-level. The can be seen as polygons with thick borders. The borders and the enclosed area have weights of opposing signs.

First of all Star builds an integral image with square and/or slanted-variants (trapezoids). The slanted ones are used to compute polygon shaped filters. Then it approximates the LoG (Laplacian of Gaussians) with choice of the bi-level polygons. The outer strip and inner strip has opposing weights. It is supposed to produce zero DC value. Meaning the ratio of the weights should somehow be related to the area under the 2 levels. Given the *m* and *n* parameters that defines of the inner region and outer boundary (thickness, sort-of). A set of filters will be defined, based on the pairs of (*m, n*) values. Repeat the convolution on the same input image with each of the scaled filter. There will be no sub sampling, therefore no need to localize points from higher-scales. Feature corners are filter-response maximas in a *3x3x3* neighbors. Eliminate weak feature points - with values below a filter response threshold. Remove unstable points on edges by examining each feature corner with Harris measure under a *9x9* window. Eliminate points having the high ratio of the two highest Principle Curvatures, above a predefined threshold.

3 Extracting Global Image Features

Despite the robustness advantages of local features, global features are still useful in applications where a rough segmentation of the object of interest is available. Global features have the ability to generalize an entire object with a single vector. Consequently, their use in standard classification techniques is straightforward. Global features include contour representations, shape descriptors, and texture features.

Global texture features and local features provide different information about the image because the support over which texture is computed varies. We expect classifiers that use global features will commit errors that differ from those of classifiers based on local features. A disadvantage of global

features is that they are sensitive to clutter and occlusion. As a result it is either assumed that an image only contains a single object, or that a good segmentation of the object from the background is available.

3.1 LBP: Local Binary Patterns

LBP [10] is a simple yet very efficient texture operator which labels the pixels of an image by thresholding the neighborhood of each pixel and considers the result as a binary number. Due to its discriminative power and computational simplicity, LBP texture operator has become a popular approach in various applications. It can be seen as a unifying approach to the traditionally divergent statistical and structural models of texture analysis. Perhaps the most important property of the LBP operator in real-world applications is its robustness to monotonic gray-scale changes caused, for example, by illumination variations. Another important property is its computational simplicity, which makes it possible to analyze images in challenging real-time settings.

The basic idea for developing the LBP operator was that two-dimensional surface textures can be described by two complementary measures: local spatial patterns and gray scale contrast. The original LBP operator forms labels for the image pixels by thresholding the 3x3 neighborhood of each pixel with the center value and considering the result as a binary number. The histogram of these 28=256 different labels can then be used as a texture descriptor. This operator used jointly with a simple local contrast measure provided very good performance in unsupervised texture segmentation. After this, many related approaches have been developed for texture and color texture segmentation.

The LBP operator was extended to use neighborhoods of different sizes. Using a circular neighborhood and bilinearly interpolating values at non-integer pixel coordinates allow any radius and number of pixels in the neighborhood.

3.2 LRF: Local Rank Functions

The experience with known features, such as Haar features and Local Binary Patterns, suggests that in many cases the classification benefits from the intensity information. On the other hand, the intensity information is subject to changes due to brightness and contrast adjustments of the images while invariance to these changes is very often wanted. This fact causes the applications using features directly based on intensity, such as Haar features, to normalize the image window being classified.

The novel Local Rank Functions (LRF) [11] is based on the idea that the intensity information in the image can be well represented by the order of the values (intensities) of the pixels or small pixel regions (e.g. summed 2×2 pixel rectangular areas). This idea is backed by the fact that calculation of

the values of features based on the order of pixels is equivalent to (or based on the exact evaluation method at least very close to) normalizing the image through histogram equalization and then evaluation of the feature value based on the pixel or small regions intensities.

The Local Rank Functions – functions based on the order of pixel values rather than the values of pixels themselves – have several principal advantages over the functions based on the values themselves:

- Invariance to illumination changes – the Local Rank Functions are invariant to most of the functions used to brightness and contrast adjustments/normalization in the images. More specifically, Local Rank Functions are invariant to nearly all monotonic gray-scale transformations.
- Strict locality – Local Rank Functions of objects (parts of objects) do not change locally when the object's image is being captured under changing conditions (similar to for example SIFT)
- Reasonable computational complexity – computation and memory accesses can be optimized thanks to regular geometric structure. No explicit normalization is needed, which is specifically important in some classification schemes.

3.3 GWT: Gabor Wavelet Transform

A Gabor filter is a linear filter used for edge detection. Frequency and orientation representations of Gabor filters are similar to those of the human visual system, and they have been found to be particularly appropriate for texture representation and discrimination. A set of Gabor filters with different frequencies and orientations may be helpful for extracting useful features from an image.

GWT is a classic method for multichannel, multi resolution analysis that represents image variations at different scales. Gabor filters are a group of wavelets obtained from the appropriate dilation and rotation of Gabor function: a Gaussian modulated sinusoid. By capturing image details at specific scales and specific orientations, Gabor filters present a good similarity with the receptive fields on the cells in the primary visual cortex of the human brain. GWT provides a flexible method for designing efficient algorithms to capture more orientation and scale information [12]. Many researches stated that GWT represents one of the efficient techniques for image texture retrieval yielding good results in content-based image retrieval applications due to many reasons:

- Being well suited for image signal expression and representation in both space and frequency domains.
- Presenting high similarity with human visual system as stated above.
- Offering the capacity for edge and straight line detection with variable orientations and scales.
- Not being sensitive to lighting conditions of the image.

3.4 HoG: Histogram of Oriented Gradients

The essential thought behind the Histogram of Oriented Gradient descriptors [13] is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The implementation of these descriptors can be achieved by dividing the image into small connected regions, called cells, and for each cell compiling a histogram of gradient directions or edge orientations for the pixels within the cell. The combination of these histograms then represents the descriptor. For improved accuracy, the local histograms can be contrast-normalized by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalize all cells within the block. This normalization results in better invariance to changes in illumination or shadowing.

The HOG descriptor maintains a few key advantages over other descriptor methods. Since the HOG descriptor operates on localized cells, the method upholds invariance to geometric and photometric transformations, except for object orientation. Such changes would only appear in larger spatial regions. Moreover, as Dalal and Triggs discovered, coarse spatial sampling, fine orientation sampling, and strong local photometric normalization permits the individual body movement of pedestrians to be ignored so long as they maintain a roughly upright position. The HOG descriptor is thus particularly suited for human detection in images.

4 References

[1] G, Lowe David, "Object recognition from local scale-invariant features," *Proceedings of the International Conference on Computer Vision*, vol. 2., pp. 1150–1157, 1999.

[2] G, Lowe David, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 6, no. 2, pp. 91-110, 2004.

[3] Bay Herbert, Ess Andreas, Tuytelaars Tinne, and Van Gool Luc, "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346-359, 2008.

[4] Edward Rosten and Tom Drummond, "Fusing points and lines for high performance tracking," *IEEE International Conference on Computer Vision*, vol. 2., pp. 1508-1511, Oct 2005.

[5] Edward Rosten and Tom Drummond, "Machine learning for high-speed corner detection," *European Conference on Computer Vision*, vol. 1., pp. 430-443, May 2006.

[6] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in *International Conference on Computer Vision*, Barcelona, 2011.

[7] Calonder Michael, Lepetit Vincent, Strecha Christoph, and Fua Pascal, "BRIEF: Binary Robust Independent Elementary Features," in *European Conference on Computer Vision*, 2010.

[8] Matas Jiri, Chum Ondrej, Urban Martin, and Pajdla Tomás, "Robust wide baseline stereo from maximally stable extremal regions," in *British Machine Vision Conference*, 2002, pp. 384-396.

[9] Agrawal Motilal, Konolige Kurt, and Rufus Blas Morten, "CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching," *Lecture Notes in Computer Science*, vol. 5305, pp. 102-115, 2008.

[10] Ojala Timo, Pietikäinen M, and Harwood D, "A Comparative Study of Texture Measures with Classification Based on Feature Distributions," *Pattern Recognition*, vol. 19, no. 3, pp. 51-59, 1996.

[11] Herout Adam et al., "Low-Level Image Features for Real-Time Object Detection," in *Pattern Recognition Recent Advances*.: InTech, 2010, pp. 111-136.

[12] Qiang Zhao, Liang Zhou, and Yin, Wuhan Chen, "Zhao Qiang ; Zhou Liang ; Chen Yin Wuhan," in *WASE International Conference on Information Engineering*, 2009, pp. 79-82.

[13] Dalal Navneet and Triggs Bill, "Histograms of Oriented Gradients for Human Detection," *International Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 886-893, 2005.