Visual Object Analysis Using Regions and Interest Points

Carles Ventura
Technical University of Catalonia
Barcelona
carles.ventura@upc.edu

ABSTRACT

This dissertation research will explore region-based and interest points based image representations, two of the most-used image models for object detection, image classification, and visual search among other applications. We will analyze the relationship between both representations with the goal of proposing a new hybrid representation that takes advantage of the strengths and overcomes the weaknesses of both approaches. More specifically, we will focus on the gPb-owt-ucm segmentation algorithm and the SIFT local features since they are the most contrasted techniques in their respective fields. Furthermore, using an object retrieval benchmark, this dissertation research will analyze three basic questions: (*i*) the usefulness of an interest points hierarchy based on a contour strength signal, (*ii*) the influence of the context on both interest points location and description, and (*iii*) the analysis of regions as spatial support for bundling interest points.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.4.6 [**Image Processing and Computer Vision**]: Segmentation; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Image Representation, Hierarchical Segmentation, Interest Points, Object Retrieval

1. MOTIVATION

Object detection, image matching, query by example, etc. are some of the applications that in the recent years have kept image processing and computer vision fields active. This dissertation research will explore two of the most used image representations: region-based and interest-points based.

Region-based image representation considers an image as a set of groups of coherent pixels, called regions, obtained by a segmentation algorithm. This way, an object is described as a set of regions and their internal properties or descriptors, which are usually based on color, texture or shape information.

Interest-points image representation is one of the most popular approaches for state-of-the-art applications, such as image classification, object detection, object recognition, and object retrieval. Interest points are detected in regions where different edges intersect as well as along edges with high curvatures. They have a spatial neighborhood whose size depends on the scale at which have been detected and over which local descriptors are computed.

The main objective of this dissertation research will be to provide a new hybrid representation that takes advantage of the strengths and overcomes the weaknesses of both approaches. This dissertation research is structured in three main parts: (*i*) Section 2 analyzes how hierarchical region-based segmentation and interest points representation can benefit one from the other, (*ii*) Section 3 focuses on exploring the influence of the context on interest points location and description, and (*iii*) Section 4 analyzes the use of regions as spatial support for bundling interest points. Each of these three parts is divided into the following subsections: (*i*) Related work, (*ii*) Proposed approach, and (*iii*) Experiments. Finally, Section 5 shows some preliminary experiments carried out.

2. HIERARCHICAL SEGMENTATIONS AND INTEREST POINTS

2.1 Related work

2.1.1 The gPb-owt-ucm segmentation algorithm

Up to now, the gPb-owt-ucm [1] is the segmentation algorithm that offers the best performance on the Berkeley Segmentation Dataset [9]. It consists of 3 different blocks: (*i*) the gPb contour detector, (*ii*) the Oriented Watershed Transform (OWT), and (*iii*) the Ultrametric Contour Map (UCM).

The gPb contour detector couples multiscale local brightness, color and texture cues to a powerful globalization framework using spectral clustering. The local cues are computed by applying oriented gradient operators at every location in the image and at different scales.

MM'13, October 21–25, 2013, Barcelona, Spain.
ACM 978-1-4503-2404-5/13/10.
http://dx.doi.org/10.1145/2502081.2502220 .

The Oriented Watershed Transform constructs the set of initial regions from an oriented contour signal. First, region minimal of a non-oriented contour signal are taken as seed locations for homogeneous segments and standard watershed transform is applied. Then, consistency between the strength of the boundaries (watershed arcs) and the underlying oriented contour signal is enforced.

The Ultrametric Contour Map is the hierarchical region tree which results from an agglomerative clustering by iteratively merging the most similar regions, i.e. the two adjacent regions which are separated by the minimum weight contour.

2.1.2 Interest points

An image can be represented as a set of interest points distributed in the image plane. A desirable property of any interest point detector is scale invariance and affine invariance. The work in [14] presents an exhaustive survey and classification of interest points detectors.

One of the first well-known detectors is the Hessian detector [2]. The Hessian matrix is defined as the matrix containing the second-order image derivatives, which encode information about how the normal to the image surface changes spatially. The trace of this matrix, known as Laplacian, is used to detect interest points. The Harris detector [3] is also widely known and it is based on detecting corner points in the image. To do so, the structure tensor or second moment matrix is computed, which describes the main directions of the gradient of the image at the neighborhood of a given location. The determinant and the trace of this matrix was proposed as a measure of cornerness of the points of the image.

The authors in [5] proposed two simple methods for extracting interest points from the segmentation maps, which focus on the boundaries and centres of the gravity of the segments.

Apart from the interest point themselves, algorithms based on local features compute local descriptors. The most known and used local descriptor is the so-called Scale Invariant Feature Transform (SIFT) [6], which is invariant to affine transformations, partially invariant to illumination changes and robust to local geometric distortion.

2.2 Proposed approach

We propose to analyze if there exists any relationship between the gPb-owt-ucm segmentation algorithm and interest points. In such a case, we would like to study if one family could benefit from the other in both directions.

First, we consider how gPb-owt-ucm could benefit from interest points. In this direction, we propose to analyze whether the interest points scales could be used to adaptively select the appropriate scale at each location. This idea emerges from [1], in which the authors detected that high resolution images of complex scenes require more than a naive weighted average of signals across the scale range. This is because such an average blurs information, resulting in good performance for medius-scale contours, but poor detection of both fine-scale and large-scale contours.

Second, we consider how interest points could take advantage of gPb-owt-ucm. In this direction, we propose to analyze whether a hierarchy of interest points could improve object retrieval performance. This idea is based on the intuition that interest points located on strong contours could be more relevant than the ones located on smoother contours. In such case, we could benefit from $gPb(x, y, \theta)$, which has rich information about contours strength at different orientations, to establish a hierarchy among the interest points. More formally, we propose two ways of weighting a keypoint:

- Without considering the keypoint orientation:

$$W(k_i) = \max_{\theta} gPb(x(k_i), y(k_i), \theta) \quad (1)$$

  where $W(k_i)$ is the weight assigned to the keypoint k_i, θ refers to the different orientations considered by gPb contour detector, and $x(k_i)$ and $y(k_i)$ are the location coordinates of the keypoint k_i. The weight is assigned taking the contour strength at the orientation which gives the maximum value.

- Considering the keypoint orientation:

$$W(k_i) = gPb(x(k_i), y(k_i), \theta(k_i)) \quad (2)$$

  where $\theta(k_i)$ is the keypoint orientation. Therefore, the weight is assigned taking the contour strenght at the orientation given by the keypoint.

Furthermore, we also propose to analyze which would be the impact on object retrieval performance of replacing interest points by the junctions detected using the gPb contour detector [7].

2.3 Experiments

With regard to the evaluation of gPb-owt-ucm benefiting from interest points, we propose using the precision-recall framework on the Berkeley Segmentation Datasets (BSDS300 and BSDS500).

On the other hand, we propose to study the object retrieval performance by using only the interest points at an increasing depth in the hierarchy. We plan to use the TRECVID benchmark dataset provided for the Instance Search task [13].

3. USE OF CONTEXT IN OBJECT RETRIEVAL AND RECOGNITION

3.1 Related work

In object recognition and retrieval, there have been suggestions that a bounding box may be able to provide some degree of context and may actually be beneficial. However, one of the contributions of [16] is the evaluation of background features, which shows the pitfalls of training on datasets with uncluttered or highly correlated backgrounds.

In [8], it is confirmed that correct spatial support is important for object recognition. Thus, knowing the right spatial support leads to substantually better recognition performance for a large number of object categories, especially those that are not well approximated by a rectangle. Overall, the recognition performance using ground-truth segments is 15% better than using the bounding boxes.

In [11], the authors went further and proposed integrating segmentation into the BoF framework considering each region as a stand-alone image by masking and zero padding the original image. Then the signature of the region is computed as in regular BoF, but discarding any feature that falls entirely outside its boundary. As a consequence, masking greatly enhances the contrast of the region boundaries making features along the boundaries more shape-informative.

Furthermore, coarse spatial information is incorporated by clustering features in segments.

3.2 Proposed approach

Considering each region as a stand-alone image by masking and zero padding the original image as in [11], the local features computed in each region do not depend on the context. Therefore, if two images contain the same object but with different backgrounds and assuming that the object is rightly segmented, the local features for both objects will be the same. However, experimental results in [11] did not estimate which was the influence of masking the region for feature extraction.

Therefore, we propose to analyze which is the influence of the context in keypoint detection. We expect that for different instances of the same object in different contexts, the locations of the keypoints will be more unstable if the object is not masked. In addition to this, context is also expected to have an impact on the description of the detected keypoints. With this goal, some measures of stability will be used to objectively measure the context influence. Furthermore, since our final application will be object retrieval, we will also measure the impact of context in object retrieval benchmarks such as the Instance Search Task of Trecvid.

On the other hand, the idea of zero padding proposed in [11] seems not to be the best way to get rid of the context. They expected that the masking and zero padding process made features along boundaries more shape-informative. However, this claim depends clearly on the color of the object. In an extreme case, no local features should be detected in an object totally black, independently of the object shape. Therefore, we think that other strategies for background should be employed to keep the shape information.

3.3 Experiments

We plan to evaluate the impact of the context in object retrieval using the TRECVID benchmark dataset provided for the Instance Search task. The goal of this task is to retrieve the videos which contain a particular object. Some instances of the object as well as their masks are given as visual examples. The task is performed for a set of objects to be retrieved. The decision of evaluate the use of context in object retrieval has also been motivated by the fact that objects to be retrieved are instance of the same query object, whereas in object recognition datasets there exist a large intra-class variance, which would blur the real impact of the context.

For both query and target images we propose the following set ups:

- Feature detection. Whether regions are masked for keypoints detection or not.
- Feature description. Whether regions are masked for keypoints description or not.

Keypoint detection and description invariance to context will be evaluated on a benchmark based on [10].

4. BUNDLING INTEREST POINTS

4.1 Related work

Classic visual vocabularies are generated from single image local descriptors. However, these vocabularies are not able to capture the rich spatial contextual information among the local features. Basically, two differents ways of considering such spatial context can be distinguish: (*i*) algorithms that apply a post geometric verification step [12], and (*ii*) algorithms which consider combination of local features [15][17]. Since we want to analyze the benefits from using regions as spatial supports for bundling interest points, we briefly describe some previous works based on combination of local features. Generally, considering visual words in groups rather than single visual word could effectively capture the spatial configuration among them.

In [15], a scheme where local features were bundled into local groups was presented based on the ideas that (*i*) each group of bundled features becomes much more discriminative than a single feature, and (*ii*) within each group simple and robust geometric constraints can be efficiently enforced. More specifically, the ellipses resulting from MSER detection are enlarged to bundle SIFT features.

In [17], a maximum number of 3 local features in each local feature group are considered since if too many local features are combined, the repeatability of the combination will decrease. Each local feature group is formed by a centered local feature and one or two other local features within a circle of radius proportional to the scale of the centered local feature. The distance between two local feature groups take into account the spatial context of each local feature group, which is defined as the orientation and scale relationships between the local features inside the group.

Recently, a dense local detector that produces repeatable shape-preserving regions via a novel segmentation-driven sampling strategy was introduced in [4].

4.2 Proposed approach

Discriminative visual phrases were proposed in [18] to refer to the frequently co-occurring visual word pairs. If two visual words frequently co-occur within short spatial distance in images containing the same object but different backgrounds, spatially consistent visual words are more likely to be located on the object.

Assuming this, we propose to work with regions resulting from a robust contour detector, such as gPb, since they can provide a good suport for bundling local features, ensuring that local features being groupped belong to a support area with minimal boundary strength. Furthermore, any boundary strength threshold on the gPb-owt-ucm generates an image partition whose regions contain internal boundaries which are smoother than the given threshold. In contrast, previous works [17][15] do not consider contour information when bundling local features to the same group.

Once adopted regions as spatial supports for bundling local features, the question about how to describe the region from the set of local features arises. Therefore, we are planning to study if support regions are enough to encode the spatial context and, therefore, classic BoF could represent the region content, or geometric relationships between the local features within the same group should be also encoded.

4.3 Experiments

We propose to evaluate the performance for different set ups which consider: (*i*) combination of local features, and (*ii*) the representation form.

With regard to the combination of local features, we propose three different cases:

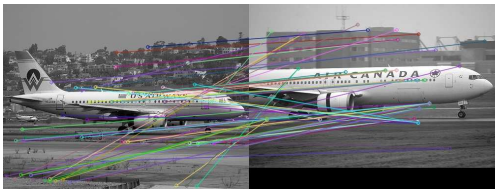

Figure 1: Matching individually SIFT descriptors

- Individual local features. This is the classic approach for both object retrieval and recognition.
- Bundling local features within the support region given by the interest point detector [17] [15].
- Bundling local features within gPb-owt-ucm regions from different hierarchical levels.

With regard to the representation form for a group of local features, the options can be mainly divided in two categories:

- No spatial information. In such case, region is consider to inherently convey enough spatial information. Classic BoF approach can be used to describe the region.
- With spatial information. Spatial coding techniques are used to enforce some geometric constraints when matching two group of local features.

5. WORK IN PROGRESS

We have carried out some preliminary experiments on using regions for bundling interest points. The images showed in Figure 1 belong to Pascal VOC 2012 Dataset. The goal of the experiment is to detect and localize the object (aeroplane) in the target image (right image of Figure 1) by using a visual example (left image of Figure 1). Keypoints and local descriptors have been obtained on each image by using the default implementation of SIFT descriptor available in OpenCV. Figure 1 shows the results of matching individually the keypoints and keeping the best matches. On the other hand, Figure 2 shows the results of region matching. For this, both query and target images have been first segmented by the gPb-owt algorithm which gives a finest partition. Then, SIFT features belonging to the each region are bundled and matched. For these preliminary experiments, the distance between two local feature groups G_1 and G_2 has been computed by averaging the distance values from each local feature of G_1 to its nearest feature of G_2.

6. ACKNOWLEDGMENTS

This dissertation research is partially founded by FPU-2010 Research Fellowship Program of the Spanish Ministry of Education.

7. REFERENCES

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2011.

[2] P. R. Beaudet. Rotationally invariant image operators. In *ICPR*, 1978.

Figure 2: Matching region-bundled SIFT descriptors

[3] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, 1988.

[4] J. Kim and K. Grauman. Boundary preserving dense local regions. In *CVPR*, 2011.

[5] P. Koniusz and K. Mikolajczyk. Segmentation based interest points and evaluation of unsupervised image segmentation methods. 2009.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[7] M. Maire, P. Arbeláez, C. Fowlkes, and J. Malik. Using contours to detect and localize junctions in natural images. In *CVPR 2008*.

[8] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentations. *BVMC*, 2007.

[9] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

[10] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI 2005*.

[11] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.

[12] J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*. 2006.

[13] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06*, 2006.

[14] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 2008.

[15] Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *CVPR*, 2009.

[16] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007.

[17] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, and Q. Tian. Building contextual visual vocabulary for large-scale image applications. In *ACM Multimedia*, 2010.

[18] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *ACM Multimedia*, 2009.