Acta Technica *62* (2017), No. 5A, 657–666

Salient object recognition based on hybrid processing model in Labelme

Junling Liu[1]

Abstract. Labelme database is used to realize the recognition of salient objects with bottom-up and top-down combination, which can deal with local features and global features simultaneously. The global gist feature information of the image describes human capture overview of structure at an instant, and the local sift features can reflect the images layout and structural relationship between these objects. The visual salient object is formed via the integration between the prior and the visual stimuli. Up-bottom prior comes from semantic information and object's regional identification information in the Labelme databases. Bottom-up visual stimuli are salient areas obtained by image processing. This method provides a new idea for visual salient object detection.

Key words. Salient object, hybrid processing, lableme.

1. Introduction

Through the research on the two modes of processing visual cognition from bottom to top and from top to bottom, we find that bottom visual stimulation can attract the allocation of resources and top visual perception and priori knowledge can well guide the detection on visual saliency [1,2]. The combination of them can raise the efficiency of detection. In a scene image, local features of image can influence the detection on visual saliency. However, global features play the leading role. If perception hybrid theory is used to simultaneously process the two features, it will contribute to rapidly achieving the detection on salient objects. Bottom-to-top and top-to-bottom modes of visual information processing are combined. local features and global features of scene images are simultaneously processed in this paper.

Feature fusion is used to detect the salient objects of scene images. Scene perception is the perception to the specific spatial relationship among several objects. The information perceived is the response of cerebral cortex to the visual stimulation of scene. The contents perceived are the effectively organized information based on top-to-bottom knowledge and experience. Therefore, scene perception is the mixed mode of processing visual information from bottom to top and from top to bottom.

[1]Workshop 1 - Jilin Engineering Normal University, Changchun,130052, China

Labelme database is used in this part. Labelme[3,4] is an image annotator based on Web and it is developed by the lab of computer science and artificial intellegence of Massachusetts Institute of Technology. This is convenient for researchers to annotate images and share with other relevant people. The database offers a matlab tool kit and many functions for application, such as checking class labels, showing the polygon of objects and counting the categories and quantity of objects. An automatic annotation tool is also included for users to define object of images and give semantics recognition labels.

When Labelme was established, the database contained 183 classes, 30369 images, 111490 labeled polygons, 44059 online editions, 67431 offline editions, 11845 static pictures and 18524 frames of sequential picture. Comparing with other databases.

A large number of quality annotations have been collected from the open data platform of Labelme, involving many different object classes. Large quantity of images in different quality, depth, width and visual angle labeled by users can be good tools for heuristic learning and training detector. The target of Labelme is to offer a dynamic data set, which is convenient for relevant research on image recognition and computer image.

2. Research plan

This part uses the hybrid scene perception and processing model proposed in Fig.1 to recognize salient objects and extract salient features. Experimental data set selects 185 images, including 1146 objects, from Labelme database 05 june 05_static _street _boston. The overall experiment plan and the concrete work process are shown in Fig.1.

Step 1 Bottom-to-top feature extraction

The global gist features of scene images are used to describe general features of images and relate them to the concentrated image class and semantics labels of Labelme.Local gift (4*4 image blocks) features of scene images are used to descibe internal objects and structure of scene images and find salient regions. Sift key feature points are used to detect and find the information of key pixels of scene images for describing details.

Step 2 Top-to-bottom priori guidance

In the phase of training, global gist features of scene image correlate to the class and semantics labels of images. Matlab tools offered by Labelme can count the number of all objects of the semantics. The labels of every scene image can be priori to find the position and contour of every object existing on image.

Step 3 Salient objects detection

Extract every object existing in scene image, regard the object as mask and combine it with the local gist features and sift feature of key points generated in the first step to detect the salientness of object.

Step 4 Extract salient features

Operate semantic classification according to the same objects existing in the same kind of images and detect the visual features relevant to the semantic objects:

Step5 Same semantic class, salient object, feature description

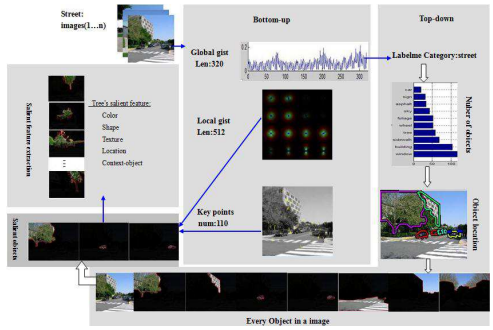

Fig. 1. Salient objects recognition model based on hybrid processing

3. Global feature and local feature extraction

Global feature of the scene is an image feature description way from global and macr image. Space envelope method [5] defines the features of the five scenes: naturalness, openness, roughness, expansion, ruggedness. Via Fourier transform, the image shall be converted to frequency domain from time domain. During this process, the structural information shall be extracted from the global features of the image. The energy spectrum after Fourier transform is expressed by high dimension of image i (x,y). During K-L transform[6], the orthogonal function set of the correlation coefficient shall be learned to resolve the random signal. Then PCA principal components shall be analyzed to achieve dimensionality reduction[7]. PCA can get uncorrelated vectors via K-L transform.

$$I(f_x, f_y) = \sum_{x=0,y=0}^{N=1} i(x,y)h(x,y)e^{-j2\pi(f_x x + f_y y)} = A(f_x, f_y)e^{j\phi(f_x, f_y)} \tag{1}$$

$$\begin{aligned} \sum_{Y} &= E\{(Y-\bar{Y})(Y-\bar{Y})^T\} = E\{T(x-\bar{x})[T(x-\bar{x})]^T\} \\ &= E\{T(x-\bar{x})(x-\bar{x})^T T^T\} = TE\{(x-\bar{x})(x-\bar{x})^T\}T^T = T\sum_{X} T^T \end{aligned} \tag{2}$$

Image I (x,y) is the two-dimensional gray-level image. It is expressed by $m \times n$ dimensional vector Γ. Thus the image training set is $\{\Gamma_i = i, \ldots P\}$. And the average vector of the image is expressed:

$$\psi = \frac{1}{P}\sum_{i=1}^{M} \Gamma_i \tag{3}$$

There exists the difference between the ith vector Γ_i and the average vector Ψ. The vector

$$\phi_i = \Gamma_i - \Psi \quad (i = 1, \ldots, P) \tag{4}$$

The covariance of the training image:

$$C = AA^T, A = [\phi_1, \phi_2, k, \phi_p] \tag{5}$$

The covariance matrix C in the formula 5 has a dimension of $m \times n$ It is kind of difficult to get the eigenvalue, so SVD singular value solution is adopted here.Theorem SVD: let A be a matrix of rank r and dimension is $n \times r$, there shall be two orthogonal matrices:

$$\begin{array}{ll} U = [u_0, u_1, \ldots, u_{r-1}] \in R^{n \times r} & U^T U = I \\ V = [v_0, v_1, \ldots, v_{r-1}] \in R^{n \times r} & V^T V = I \\ \Lambda = diag[\lambda_0, \lambda_1 \ldots, \lambda_r] \in R^{n \times r} & \end{array} \tag{6}$$

In the formula 6, Λ is a diagonal matrix, and $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_r$

$$A = U\Lambda^{\frac{1}{2}}V^T \tag{7}$$

In the formula 7, $\lambda_i (i = 0, 1, \ldots, r-1)$ is the nonzero eigenvalue of matrix $AA^T ? A^T A$ and their eigenvectors are u_i and v_i respectively. $\sqrt{\lambda_i}$ is the singular value of A.

Deduction:

$$U = AVA^{-\frac{1}{2}} \tag{8}$$

Based on SVD solution, the eigenvalue and eigenvector of matrix C can be solved to get the eigenvalue and eigenvector of its transposed matrix L. Since L is $P \times P$, let the singular value of L be $v_i (i = 1, 2, \ldots, P)$

$$L = A^T A \tag{9}$$

The eigenvector of matrix C $u_i i = 1, \ldots, P$ can be solved by difference image $\Phi_i(1, \ldots, P), u_i(1, \ldots, P)$ linear combination:

$$U = [u_1, u_2, \ldots, u_P] = [\Phi_1, \Phi_2, \ldots \Phi_P][\nu_1, \nu_1, \ldots \nu_P] = AV \tag{10}$$

Smoothing can be used to divide the image into 4*4 areas (equal but not overlapping). Later Fourier transform can be used to extract the global features of the image, like expansion, roughness etc. The global feature generally adopts the frequency domain energy spectrum to calculate representation. Finally PCA is used for dimensionality reduction to get Gist feature information to classify the scene images. Thirty-four subchannels (direction: $0°, 45°, 90°, 135°$, four scales, 16 subchannels in total.color: red-green, blue-yellow surround 6 scales and 12 subchannels; brightness: darkness-lightness surround 6 scales and 6 subchannels) are used for Gist feature extraction. Sixteen features shall form 544 dimension. PCA dimensionality reduction can get 320 features.

In 2004, Lowe put forward sift algorithm [8], which made it possible for the image to extract the invariant feature with scale, rotation, affine and space perspective transform. The scale space of image I(x,y) under the many scales can be seen in the formula 11. (x,y) represents the pixel location of the image, and is the scale space

factor. The smaller is, the smaller the corresponding scale is. The image of small scale can better reflect the image details. The image of large scale can better express the image overview.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{11}$$

DoG is defined as the difference of gauss kernel of the two different scales:

$$D(x, y,) = (G(x, y, k) - G(x, y,)) * I\,(x, y)\ = L(x, y, k\sigma) - L(x, y, \sigma) \tag{12}$$

To guarantee the rotation invariance of the operator, the gradient direction distribution of the key point neighborhood pixel shall be adopted. The direction distribution features shall provide the direction parameters for each key point.

$$\begin{array}{l} m(x, y) = \sqrt{(L(x+1), y) - L(x-1, y))^2 + ((L(x), y+1) - L(x, y-1))^2)} \\ \theta(x, y) = atan2((L(x, y+1) - L(x, y-1))/(L(x+1, y) - L(x-1, y))) \end{array} \tag{13}$$

The module value m and the direction θ of the image point (x,y) gradient are calculated in formula 13. L represents the scale of (x,y) key point. The calculation sampling is carried out in (x,y) neighborhood window. The histogram computation method can be used to get the gradient direction of the neighborhood pixel. The value range of the histogram gradient direction is $0 \sim 360°$. Compressed sampling 10 degrees is one column, and there are 36 columns. The maximum direction of the statistical histogram represents the main direction of the key point (x,y) neighborhood gradient as the direction of the key point.

One key point consists of four seed points. Each seed point has eight direction vectors. Such neighborhood multiple-seed and multiple-direction description method makes sift algorithm full of noise immunity and fault tolerance. At the moment, SIFT feature vector is free from the influence of scale change, rotation and other geometry deformation factors. If the feature vector length is normalized further, the illumination variation influence[9] shall be removed.

4. Salient object detection

The salient objects to be detected are the objects labeled in the scene. Their local gist features are salient and several sift key points exist simultaneously. As for extracting the regions with salient local gist features, the pixel G(x,y) is 1 if it is salient. otherwise, G(x,y) is 0 (Fig.2 (a)). As for the feature points extracted by sift, the larger norm of eigenvector, the more salient features of key points (Fig.2 (b)). The position and region of objects in scene images existing in Labelme database are shown in Fig.2 (c). The process that multi-feature fusion is used to form salient objects is shown in Fig.2.

Find the regions with salient visual features according to bottom-to-top Fig.2 (a) and Fig.2 (b) and compare them with the priori knowledge of objects offered by Labelme Fig.2 (c). If the result of match between key feature points and the

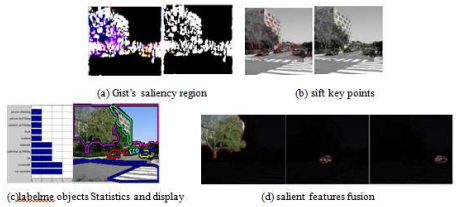

Fig. 2. The salient object detection

pixel points existing in objects is larger than threshold value and the salientness of gist region exists, the region is salient object. The contribution degree of every key point of sift is considered at the same time. The gradient information in sift is the eigenvector of key point; norm of vector reflects the size of gradient; the key points with large gradient have visual salientness in high degree. Therefore, the larger norm is, the higher contribution degree will be. Details are explained in formula 14:

$$\frac{\sum_{i=1}^{Onum}\sum_{z=1}^{Snum}|sift\_l|}{\sum_{i=1}^{Onum}\sum_{j=1}^{Pnum}I(x,y)\times G(x,y)} \geq threshold \tag{14}$$

In the formula: Onum is the number of certain objects; Pnum is the number of pixels in certain object; Snum is the number of sift key points of certain object; Sift_l is the vector of sift key points; |sift_l|is the norm value of vector; I(x,y) is the gray value after pixels are normalized; threshold is average threshold value.

5. Extract salient features

Operate semantic classification according to the same objects existing in the same kind of images and detect the visual features relevant to the semantic objects.Transform the brightness of RGB colorful images[10]:

$$Lu\min ance(x,y) = r(x,y)\times 0.299 + g(x,y)\times 0.587 + b(x,y)\times 0.114 \tag{15}$$

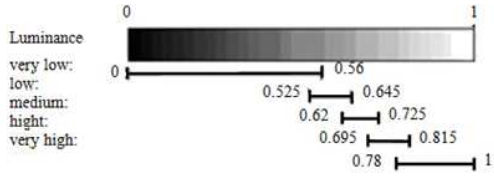

Fig. 3. The quantified luminance feature

To better recognize and match the salient objects in scene perception process, Fig.3 according to extracting the visual features of objects based on Labelme polygon. This method is simple and easy and accords with humans' perception character. At the same time, this part puts forward the method of quantifying and extract-

ing other features including L-a-b color space, size and texture and simplifies the description of feature space.

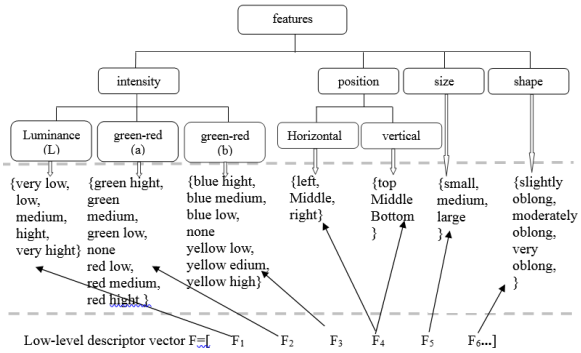

Fig. 4. The quantified visual feature in low-level

The quantized features of brightness and position can better meet the visual perception of humans. Quantized feature description can simplify the data set of knowledge base, which contributes to recognizing targets.

The quantized features of brightness and position can better meet the visual perception of humans. Quantized feature description can simplify the data set of knowledge base, which contributes to recognizing targets.

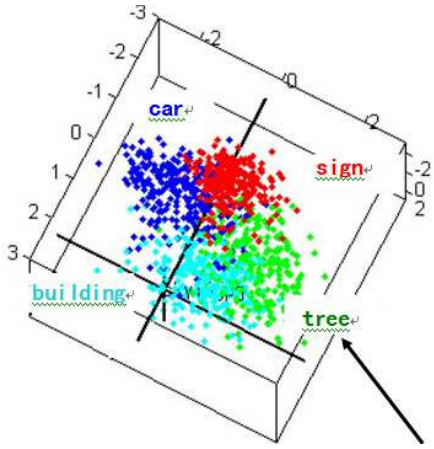

Fig. 5. Salient objects feature space

6. Results and discussion

This part researches the two processing modes of visual cognition from bottom to top and from top to bottom. we find that bottom visual stimulation can attract the allocation of resources and top visual perception and priori knowledge can well guide the detection on visual saliency. The combination of them can raise the efficiency of

detection. In a scene image, local features of image can influence the detection on visual saliency. However, global features play the leading role. If perception mixture theory is used to simultaneously process the two features, it will contribute to rapidly achieving the detection on salient objects. Bottom-to-top and top-to-bottom modes of visual information processing are combined; local features and global features of scene images are simultaneously processed in this paper. Feature fusion is used to detect the salient objects. Global features describe the general information of images got by humans while instantaneously scanning them. Local features can reflect internal layout of images and the structural relation among different targets. Pixel feature points can better describe the detailed information of images for matching and tracing. This part extracts the global gist information and local gist information of scene images. At the same time, sift description operators are used to describe the internal key points of target. Global and local information guides visual sense to form the region with salient features. The labeling information of Labelme is used; semantic information and the regional sign information of objects are regarded as apriori information to combine with salient regions at bottom layer to form visual salient targets.

References

[1] A. Borji: *Boosting bottom-up and top-down visual features for saliency estimation.* CVPR (2012), 438–445.

[2] Y. Yuan, D. Li, M. Q. Meng: *Automatic Polyp Detection via A Novel Unified Bottom-up and Top-down Saliency Approach.* IEEE Journal of Biomedical and Health Informatics *99* (2017), 1–8.

[3] A. Torralba, R. Fergus, W. T. Freeman: *80 Million tiny images:a large dataset for non-parametric object and scene recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI) *30* (2008), 1958–1970.

[4] B. C. Russell, A. Torralba, K. P. Murphy, P. LabelMe: *a database and web-based tool for image annotation.* International journal of computer vision *77* (2008), 157–173.

[5] A. Oliva: *Antonio Torralba.Modeling the shape of the scene: a holistic representation of the spatial envelope.* International Journal of Computer Vision *42* (2001), 145–175.

[6] K. Zhou, J. Cai, Y. H. Xu: *Osteoporosis Recognition Based on Similarity Metric with SVM.* International Journal Bioautomation *20* (2016) 253–264.

[7] D. G. Lowe: *Distinctive image features from scale-invariant keypoints.* International Journal of Computer Vision *60* (2014), 91–110.

[8] Z. Luo, Y. Jia: *MR Image Contrast Enhancement by Wavelet-based Contourlet Transform.* International Journal Bioautomation *20* (2016), 265–278.

[9] P. A. Koringa, S. K. Mitra, V. K. Asari: *Handling Illumination Variation: A Challenge for Face Recognition.* Proceedings of International Conference on Computer Vision and Image Processing (2017), 273–283.

[10] B. T. Grys, D. S. Lo, N. Sahin: *Machine learning and computer vision approaches for phenotypic profiling.* J Cell Biol *216*, (2016), 1–7.

Received November 16, 2017