Research on the Contextual Information in Scene Classification

Pan Feng, Danyang Qin(✉), Ping Ji, and Jingya Ma

Key Lab of Electronic and Communication Engineering,
Heilongjiang University,
No. 74 Xuefu Road, Harbin, People's Republic of China
qindanyang@hlju.edu.cn

Abstract. The classical localization approaches only focus on the performance of features extracted from images but ignore contextual information hidden in the images. In this paper, it is annotated on the images and SVM model is used to classify different images for semantic localization. Supervised Latent Dirichlet Allocation (sLDA) model is introduced to obtain the annotations, and the standard SIFT algorithm is improved to extract feature descriptors. Two situations are designed for the acquisition of contextual annotations, which are to provide the accurate contextual annotations directly and to infer contextual information by sLDA model. The effect of contextual information in scene classification is simulated and verified.

Keywords: Contextual information · Semantic localization
Scene classification

1 Introduction

With the development of Artificial Intelligence (AI), the robot localization has become a research hotspot. Considering the importance of robot localization, some existing localization methods are combined to obtain better performance.

The semantics-based visual localization method is adopted in this paper, which takes use of the category labels such as "office" and "corridor" to describe the location of the robot and can be applied to many cases. [1] proposed an application of semantic localization to autopilot. As the camera is the primary information collection device of the robot, the semantic positioning can be considered as a classification problem.

To improve the precision of classification, contextual information annotations are combined with image feature descriptors. Contextual information involves multiple aspects e.g. keywords to describe images. In other fields, some papers are proposed to solve various technology problems with contextual information. Contextual

This work was supported by the National Natural Science Foundation of China (61771186), Postdoctoral Research Project of Heilongjiang Province (LBH-Q15121), University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province (UNPYSCT-2017125).

L. Meng and Y. Zhang (Eds.): MLICOM 2018, LNICST 251, pp. 338–345, 2018.
https://doi.org/10.1007/978-3-030-00557-3_34

information can be used to monitor system intrusion from external network [2]. Filippini in [3] adopt the contextual information of user location to improve directional cell discovery.

In addition to the contextual information, the standard Scale Invariant Feature Transform (SIFT) method will also be adopted to improve the performance of image descriptors, and the sLDA model [4] is used to obtain the contextual annotations.

Three scenes will be designed and adopted to perform the simulation: (a) without contextual information; (b) providing context information annotation directly; (c) inferring contextual information by sLDA. Scene (a) represents the classic semantic localization, scene (b) will show the maximum benefit of the contextual information and scene (c) will evaluate the effective integration of contextual information.

2 Related Work

2.1 Image Feature Descriptor

Image feature descriptors are used to describe image features, such as color, shape and gradient, which are divided into local feature descriptor and global feature descriptor. In our work, we use two feature descriptors, Histogram of Oriented Gradient (HOG) [5] and Histogram of Vision Words (HoVW), which is combination of the SIFT and the Bag of Words (BoW).

The BoW process is based on cluster and takes use of the local features extracted from image as the input. Every cluster is a "word", and the whole n words are defined as a whole to be a codebook. By mapping local features to words, any input image can be represented as a word bag. Finally, the word frequency histogram is calculated and it is a global feature descriptor.

In addition, different descriptor dimensions will be evaluated in the following ways: (a) Combining neighboring angles in HOG process; (b) Choosing different numbers of words in HoVW process. Finally, four dimensions of 50, 100, 200 and 300 are selected and shared by two descriptors.

2.2 sLDA Model

LDA is a kind of Bayesian model, which can infer the posterior distribution of hidden variables as in (1) if given a set of visual words in the image.

$$P\left(\theta_d, z_j | w_i, \alpha, \beta\right) = \frac{P\left(\theta_d, w_i, z_j | \alpha, \beta\right)}{P(w_i | \alpha, \beta)} \tag{1}$$

where z_j is the topic of the visual word w_i that can be observed, and is defined by $P\left(z_j | \theta_d\right)$. $Z = \{z_1, z_2, \ldots, z_k\}$, $|Z| = k$ indicating that there are k topics. θ_d is the mixture proportion of the topics in the image and it is a Dirichlet random variable. If there is z_j, w_i can be obtained from $P\left(w_i | z_j, \beta\right)$ under multinomial distribution, where β is a $k \times V$ matrix and $\beta_{i,j} = P(w^j = 1 | z^i = 1)$ as in (2).

$$P(w_i|z_j,\beta) = \prod_{n=1}^{N} \beta_{z_j,w_i} \tag{2}$$

w_i is defined as (3):

$$P(w_i|\alpha,\beta) = \int P(\theta_d|\alpha)\left(\sum_{j=1}^{k} P(z_j|\theta_d)P(w_i|z_j,\beta)\right)\mathrm{d}\theta_d \tag{3}$$

where α and β are model parameters defined before test. $P(\theta_d|\alpha)$ is defined as (4):

$$P(\theta_d|\alpha) = \frac{\Gamma(m\alpha)}{\Gamma^m(\alpha)} \prod_{j=1}^{m} \theta_{d_j}^{\alpha-1} \tag{4}$$

While sLDA adds a response variable y to LDA and jointly model the document and the response to find latent topics which can predict the response variables for unlabeled images in the future.

The response variable comes from a normal linear model $N(\eta^{\mathrm{T}}\bar{z}, \sigma^2)$, where η and σ are the response parameters and there is $\bar{z} = 1/N \sum_{j=1}^{N} z_j$.

A graphical model representation of sLDA can be seen in Fig. 1. Top-N F-measure in [6] is used to measure annotation performance and there is $N = 5$. The score is standardized to represent a number between 0 and 1, where the larger the number is, the stronger the relevancy will be.

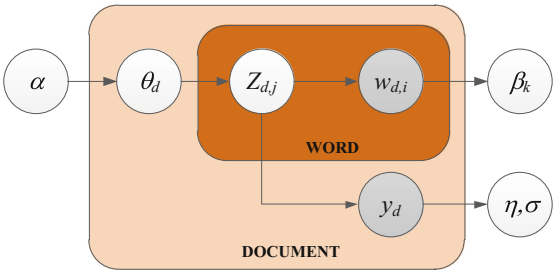

Fig. 1. Graphical model representation of sLDA

Table 1 shows performance comparison of sLDA and multi-label SVM in object recognition, which indicates the performance of sLDA is better than that of SVM.

2.3 SIFT Algorithm Optimization

Considering that SIFT cannot deal with images in many scenes well enough, such as underwater scene. An optimized SIFT is presented in this section.

Table 1. Object detection performance of sLDA and SVM

Detecting object	SVM	sLDA
Bed	0.5371	0.6322
Cupboard	0.3029	0.4204
Keyboard	0.5157	0.6846
Monitor	0.4529	0.6212
Table	0.3829	0.4843

Pre-filtering operation in the image by a Gabor filter can be realized as follows:

$$g_{x,y,\theta} = \frac{1}{2\pi\sigma_1\sigma_2} \times \exp\left(i\frac{2\pi}{\lambda}(x\cos\theta + y\sin\theta)\right) \times \exp\left(-\frac{(x\cos\theta + y\sin\theta)^2}{2\sigma_1^2} - \frac{(y\cos\theta - x\sin\theta)^2}{2\sigma_2^2}\right) \tag{5}$$

where (x, y) is the coordinate; θ is the orientation of the filter; λ is the wavelength; σ_1 and σ_2 are Gaussian standard deviations taken along with the orientation θ and $\theta + \pi/2$.

To approximate the odd Gabor filter, there is $\sigma_1 = \sigma_2 = \sigma$ to make only one variable exist in the function. Odd Gabor filters approximate odd Gaussian filters where $\lambda = 6\sigma$, so it can be set as $\lambda = 6\sigma$ and being taken into (5) to generate (6). It is similar to an odd Gaussian filter.

$$g_{x,y,\theta} = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)\sin\left(\frac{2\pi}{6\sigma}(x\cos\theta + y\sin\theta)\right) \tag{6}$$

As standard SIFT has a fixed threshold as 0.03, many key points in dim scenes will be eliminated. To avoid the absence of information, the threshold will be kept at 10% of the image contrast. The points below 10% are considered as the low illuminance points and are not regarded as key points.

Standard SIFT uses pixel differentiation to obtain the image derivative and further generates relative amplitude and gradient. The process will be very sensitive to noise. The pixel difference process involves high-pass filtering, amplifying high-frequency noise in it. To avoid the noise, the sobel operator is adopted to each key point.

$$\begin{aligned} M(x,y) &= \sqrt{D_x(x,y)^2 + D_y(x,y)^2} \\ \Theta(x,y) &= \tan^{-1}\left(D_y(x,y)/D_x(x,y)\right) \end{aligned} \tag{7}$$

where the intermediate variables satisfy:

$$\begin{aligned} D_x(x,y) &= f_{x_{sobel}}(x,y)I(x,y) \\ D_y(x,y) &= f_{y_{sobel}}(x,y)I(x,y) \end{aligned} \tag{8}$$

and

$$f_{x_{sobel}} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, f_{y_{sobel}} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{9}$$

The parameters $f_{x_{sobel}}$ and $f_{y_{sobel}}$ are the sobel operators along the orientation x and y, and $M(x, y)$ and $\Theta(x, y)$ are value and orientation of gradient respectively. This improvement will preserve more information in the descriptors and eliminate noise.

Finally, we use Hausdorff distance [7] to computer the distance between key points. Given two sets of points $A = \{a_1, a_2, \ldots, a_m\}$ and $B = \{b_1, b_2, \ldots, b_n\}$, Hausdorff distance is calculated as follows:

$$H(A, B) = \max(h(A, B), h(B, A)) \tag{10}$$

where $h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|$. Hausdorff distance is more accurate than Euclidean distance between two key points and eliminates false matches. The performance comparison of SIFT before and after optimization is shown in Table 2, in which Key Points in Reference Image (KPRI), Key Points in Test Images (KPTI), Match points (Matches), Correct Match points (CMathes), Root Mean Square Error (RMSE) and Time Cost will be taken to compare based on standard and optimized SIFT.

Table 2. Comparison of quantitative parameters

Algorithm	KPRI	KPTI	Matches	CMathes	RMSE	Time Cost
SIFT (standard)	259	493	44	4	1.79	5.41
SIFT (optimized)	1464	2034	32	5	0.69	15.1

2.4 SVM Model

Supposing A is a series of examples and labels under unknown probability distributions, it needs to find a function that allows the most accurate determination of the class of any future example. Generally, there is:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i K(x, x_i) + b \tag{11}$$

where $b \in R$, b and α_i are Lagrange coefficients. Most α_i will become zero after training and the vector with non-zero is called support vector. $K(x_1, x_2)$ is the kernel function being selected based on specific issue.

Most classification models can solve multi-label problem, but SVM model fail to do so. Multi-label SVM classifier can be constructed by using one-versus-one or one-versus-all method [8], and the one-versus-all is adopted in the test. About the key parameter $K(x_1, x_2)$, the following two metrics are usually taken to evaluate:

- Linear kernel (SVM_{lin}): $K(x_1, x_2) = x_1 \cdot x_2 + coef_0$
- χ^2 kernel (SVM_{χ^2}): $K(x_1, x_2) = 1 - \sum_{i=1}^{n} \frac{(x_1 - x_2)^2}{\frac{1}{2}(x_1 + x_2)}$

Considering about the indoor characteristics, χ^2 kernel is taken in this paper. The paper evaluates the model through 5-fold cross validation. To keep each sample distribution invariable in the test, we use stratified fold selection. In addition, the fold remains the same value during evaluation of different descriptors, and the effect of randomness is avoided.

2.5 Context Information

Two existing datasets are adopted to train and test the model as KTH-IDOL2 [9] and ViDRILO [10], both of which are acquired by robots in indoor environments:

- KTH-IDOL2 dataset contains 5 scenes and 3 lighting conditions
- ViDRILO dataset contains 10 scenes and the existence of 15 objects in images

The lighting condition in KTH-IDOL2 and the existence of 15 objects in ViDRILO are considered as contextual information respectively. By adding some binary numbers to descriptors, we combine context information with descriptors.

Contextual information about the objects is not exclusive and each object is a binary variable. Although the lighting conditions are unique, we choose 3 binary representations allowing more experimental variables, and 15 binary values are taken to represent objects' presences which are annotated in the ViDRLO.

3 Performance Evaluation

We choose HoVW and HOG as feature descriptors, and the dimensionalities are chosen for 50, 100, 200 and 300. 5-fold cross validation is used to calculate classification accuracy. In the test, image descriptors are considered to be input data of model to obtain context information.

Figure 2 shows the simulation results in three cases: (1) Free of contextual information (Baseline); (2) Providing context annotations (Ideal) directly; (3) Inferring contextual information (Realistic). Comparison and analysis from Fig. 2 can draw the following conclusions:

- Annotations in KTH-IDOL2 have no effect on scene classification, suggesting that the lighting condition has little to do with the scene category;
- Comparing HoVW and HOG, the combination with lower baseline accuracy has larger improvement space by integrating contextual information;
- When inferring contextual information, the SVM classification is less effective than scene without contextual annotations, indicating SVM is sensitive to data error;

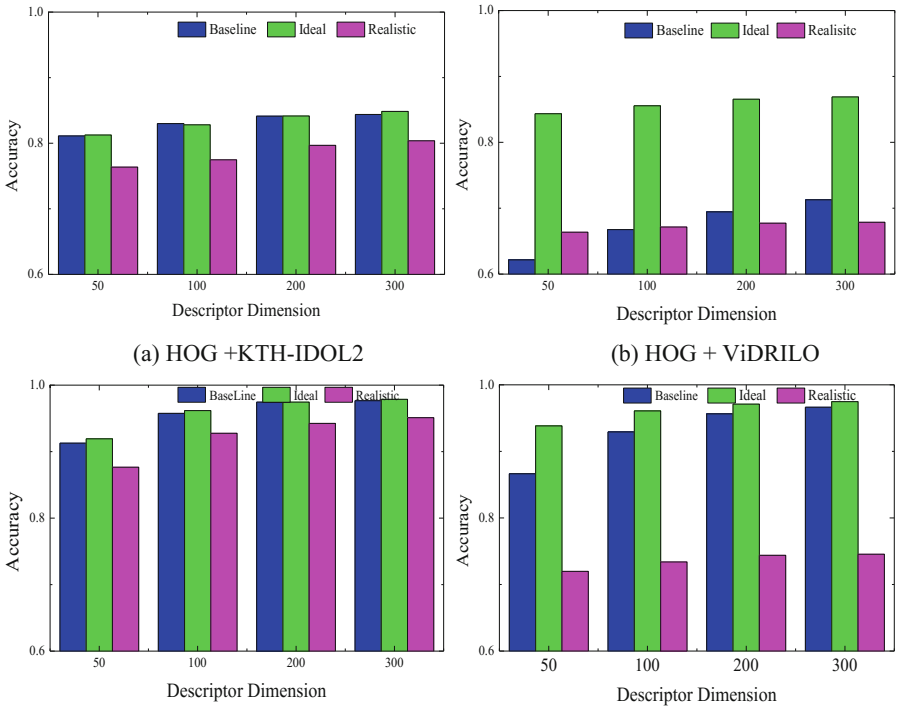

Fig. 2. Simulating results of scene classification tests

4 Conclusion

To achieve the scene classification effectively in real application, the combination of contextual information and image descriptor are proposed and evaluated in this paper. Two integration methods are adopted: one is giving accurate contextual information annotations and the other is inferring contextual information at the initial stage. The proposal is experimentally tested by using SVM classification model with two datasets and two image descriptors. It can be concluded that the contextual information is useful for classification.

The effect of context information relies on the descriptor, model and dataset. When providing contextual information directly, the classification accuracy improves; when inferring contextual information, there are some errors which make the classification result worse. In the future, we will do more experiments on new datasets, classification models and descriptors to find more effective approaches using contextual information.

References

1. Sefati, M., Daum, M., Sondermann, B., Kreisköther, K.D., Kampker, A.: Real-time vision-aided localization and navigation based on three-view geometry. In: International Conference on Intelligent Vehicles, Los Angeles, CA, USA, pp. 13–19. IEEE (2017)
2. Anton, S.D., Fraunholz, D., Schotten, H.D., Teuber, S.: A question of context: enhancing intrusion detection by providing context information. In: International Conference on Internet of Things Business Models, Users, and Networks, Copenhagen, Denmark, pp. 1–8. IEEE (2017)
3. Filippini, I., Sciancalepore, V., Devoti, F., et al.: Fast cell discovery in mm-wave 5G networks with context information. IEEE Trans. Mob. Comput. **99**, 1 (2017)
4. Blei, D.M., Mcauliffe, J.D.: Supervised topic models. Adv. Neural. Inf. Process. Syst. **3**, 327–332 (2010)
5. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: International Conference on Image and Video Retrieval, The Netherlands, Amsterdam, pp. 401–408. ACM (2007)
6. Wang, C., Blei, D., Li, F.F.: Simultaneous image classification and annotation. In: Computer Vision and Pattern Recognition, pp. 1903–1910 (2010)
7. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. **15**(9), 850–863 (1993)
8. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. J. Mach. Learn. Res. **5**, 101–141 (2004)
9. Luo, J., Pronobis, A., Caputo, B., et al.: The KTH-IDOL2 database (2006)
10. Martinez-Gomez, J., Cazorla, M., Garcia Varea, I., et al.: ViDRILO: the visual and depth robot indoor localization with objects information dataset. Int. J. Robot. Res. **34**(14), 1681–1687 (2015)