Research Journal of Applied Sciences, Engineering and Technology 7(5): 970-974, 2014
DOI:10.19026/rjaset.7.343
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 29, 2013 Accepted: February 25, 2013 Published: February 05, 2014

Research Article

Scene Semantics Recognition Based on Target Detection and Fuzzy Reasoning

[1]Weiliang Liu, [2]Changliang Liu and [1]Yongjun Lin
[1]Department of Automation, North China Electric Power University, Baoding, China
[2]State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources,
Beijing, China

Abstract: In order to get better image semantics recognition, a recognition system based on object detection and fuzzy reasoning is presented in this study. The system contains four parts: image preprocessing, feature extraction, target recognition and fuzzy reasoning machine. Compared with other methods, the outputs of target detectors are fuzzed, the fuzzy relationships between targets are extracted and fuzzy inference is performed using fuzzy automata. The experiment indicates that this method could overcome the problems of false positive and false negative of pattern classifiers and perform relatively more accurate image semantics recognition than other existing methods.

Keywords: Feature extraction, fuzzy reasoning machine, image preprocessing, semantics recognition, target detection

INTRODUCTION

As one kind of important information resources, image contains far more knowledge than words. Content-Based Image Retrieval (CBIR) has many advantages compared with other image retrieval method and becomes the mainstream of image retrieval technology. However, CBIR can't support semantics retrieval due to it mainly adopts the bottom image visual features such as color, texture, shape, etc. Semantics-based image retrieval technology is considered as a more ideal and valuable image retrieval method, therefore, how to identify image semantics well using the knowledge reasoning system has become a research hotspot.

Image Semantics refers the meaning of the scene or behavior in image, which consists of targets and their relations. In order to get image semantics, many specific image processing and recognition problems must be solved such as image segmentation, reconstruction, edge detection, feature extraction, target detection, etc. In recent years image processing and detection technology have made great progress, such as iris identification (Boles and Boashah, 1998), face recognition (Viola, 2001) and so on. But for some more complex target identification, such as severe non-rigid target detection, is still has great difficulties (Li and Hu, 2005). Therefore, research work of how to perform relatively precise semantics recognition based on inaccurate target detection is very meaningful.

Due to the uncertainty of existing target detection technology, detection result could be expressed as fuzzy information, which the fuzzy function could deal with. For better realization of image semantics recognition, this study proposes a scene semantics identification system based on fuzzy reasoning. The system consists of the following four parts: image preprocessing, feature extraction, target detection and fuzzy reasoning machine. Figure 1 gives the system framework of the presented scene semantics recognition. Compared with other methods, the outputs of target detectors are expressed by fuzzy information, fuzzy relationships between targets are extracted and fuzzy reasoning is performed using fuzzy automata. The simulation indicates that this method could overcome the problems of false positive and false negative of target detectors and perform relatively more accurate image semantics recognition than other existed methods.

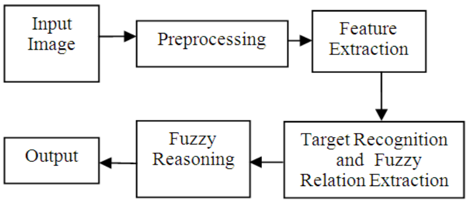

Fig. 1: Flowchart of Image semantics recognition system

Image preprocessing: The aim of image preprocessing is to perform better feature extraction and target detection. The image collected by record equipment includes not only the targets, but also some noise. With the influence of light or other reasons, the image may

Corresponding Author: Weiliang Liu, Department of Automation, North China Electric Power University, Baoding 071003, China

be vague, which makes the next feature extraction and exact matching difficult. Image preprocessing contains image noise reduction, image smooth, image enhancement, etc. Here existing image processing technologies are adopted, such as the two-dimensional wavelet transform, Gabor filters, etc.

FEATURE EXTRACTION

The aim of feature extraction is to get a group of "fewer but better" classification feature. Because the generation of feature is associated with specific target, there lack of a unified feature extraction method which is suitable for various targets in theory. Therefore, for different target, different feature extraction methods are adopted here. For face, Haar wavelet extraction method is adopted (Viola, 2001) For human body, gradient directed graph characteristics (HOG) is adopted (Dalal and Triggs, 2005; Dalal, 2006); For some background such as grassland, sea, beach, etc, using color and texture features may be a good choice. This section introduces the Gabor filter, which is widely used in target detections like fingerprint detection, iris recognition and so on.

2-D Gabor function can be expressed as:

$$g\left(x,y\right)=\frac{1}{2\pi\delta_x\delta_y}\cdot e^{-\frac{1}{2}\left[\left(\frac{x}{\delta_x}\right)^2+\left(\frac{y}{\delta_y}\right)^2\right]+j\left(ux+vy\right)} \tag{1}$$

And its Fourier Transform is:

$$\hat{g}\left(\omega_x,\omega_y\right)=e^{-2\pi^2\left[\delta_x^2\left(\omega_x-\frac{u}{2\pi}\right)^2+\delta_y^2\left(\omega_y-\frac{v}{2\pi}\right)^2\right]} \tag{2}$$

The two-dimensional Gabor function is the result of translation of a two-dimensional Gaussian function at the two frequency axes. The two-dimensional Gaussian function is a two-dimensional smoothness function, while the two-dimensional Gabor function is a two-dimensional band-pass filter. Performing convolution on Gabor function on *g* (*x, y*) and 2-D texture image *p* (*x, y*), that is:

$$f\left(x,y\right)=g\left(x,y\right)\otimes p\left(x,y\right) \tag{3}$$

Good result of image edges could be obtained. Usually, the Gabor function is designed as:

$$g\left(x,y\right)=e^{-\frac{1}{2}\left[\left(\frac{x}{16}\right)^2+\left(\frac{y}{16}\right)^2\right]+j\left(\frac{\pi}{8}x+\frac{\pi}{8}y\right)} \tag{4}$$

Assume the model of every channel is:

$$\begin{cases} q\left(x,y\right)=\sqrt{q_e^2\left(x,y\right)+q_o^2\left(x,y\right)} \\ q_e\left(x,y\right)=h_e\left(x,y\right)\otimes p\left(x,y\right) \\ q_o\left(x,y\right)=h_o\left(x,y\right)\otimes p\left(x,y\right) \end{cases} \tag{5}$$

where *p* (*x, y*) is the input image of the channel, h_e (*x, y*), h_o (*x, y*) are respectively the even Gabor filter and odd Gabor filter. Without losing of universality, isotropy Gabor filter can be used (Tan, 1995):

$$\begin{cases} h_e\left(x,y,f,\theta,\sigma\right)=g\left(x,y,\sigma\right)\cdot\cos\left[2\pi f\left(x\cos\theta+y\sin\theta\right)\right] \\ h_o\left(x,y,f,\theta,\sigma\right)=g\left(x,y,\sigma\right)\cdot\sin\left[2\pi f\left(x\cos\theta+y\sin\theta\right)\right] \end{cases} \tag{6}$$

where *g*(*x*, *y*, σ) is the Gaussian function:

$$g\left(x,y,\sigma\right)=\frac{1}{2\pi\sigma^2}\cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{7}$$

f, Ɵ, σ in (6) and (7) are respectively the important parameters of Gabor filter, i.e., spatial frequency, phase and spatial constant. The Fourier Transform of even Gabor filter and odd Gabor filter in Eq. (6) are:

$$\begin{cases} H_e\left(\omega_x,\omega_y,f,\theta,\sigma\right)=\frac{\left[H_1\left(\omega_x,\omega_y,f,\theta,\sigma\right)+H_2\left(\omega_x,\omega_y,f,\theta,\sigma\right)\right]}{2} \\ H_o\left(\omega_x,\omega_y,f,\theta,\sigma\right)=\frac{\left[H_1\left(\omega_x,\omega_y,f,\theta,\sigma\right)-H_2\left(\omega_x,\omega_y,f,\theta,\sigma\right)\right]}{2j} \end{cases} \tag{8}$$

and

$$\begin{cases} H_1\left(\omega_x,\omega_y,f,\theta,\sigma\right)=e^{-2\pi^2\sigma^2\left[\left(\omega_x-f\cos\theta\right)^2+\left(\omega_y-f\sin\theta\right)^2\right]} \\ H_2\left(\omega_x,\omega_y,f,\theta,\sigma\right)=e^{-2\pi^2\sigma^2\left[\left(\omega_x+f\cos\theta\right)^2+\left(\omega_y+f\sin\theta\right)^2\right]} \end{cases} \tag{9}$$

In practice, convolution is usually performed using Fourier Transform, that is:

$$\begin{cases} q_e\left(x,y\right)=h_e\left(x,y\right)\otimes p\left(x,y\right)=FFT^{-1}\left[P\left(\omega_x,\omega_y\right)\cdot H_e\left(\omega_x,\omega_y,f,\theta,\sigma\right)\right] \\ q_o\left(x,y\right)=h_o\left(x,y\right)\otimes p\left(x,y\right)=FFT^{-1}\left[P\left(\omega_x,\omega_y\right)\cdot H_o\left(\omega_x,\omega_y,f,\theta,\sigma\right)\right] \end{cases} \tag{10}$$

where, $P\left(\omega_x,\omega_y\right) = [p(x,y)]$.

Every pair of Gabor filters, h_e(x, y), h_o(x, y), correspond to given spatial frequency and direction. The frequency information and direction information could be extracted at the same time. As to the aim of texture recognition, it is unnecessary to choose the filter parameter space that covers the whole frequency field (Tan, 1995). The smaller the center frequency is, the larger scale of extracted texture feature is. Usually the frequency which is the exponent of 2 is chosen. For example, in the fingerprints recognition algorithm, we choose 2, 4, 8, 16, 32, 64 as center frequency respectively. For every center frequency, four phases are selected: $\theta = \frac{\pi}{4}, \frac{\pi}{2}, \frac{3}{4}\pi$. In this way, there are 24 channels of Gabor filter. To the filter result of every channel, the exception and deviation are extracted as its features. To every input image, multicenter Gabor filter could extract 48 features in total.

TARGET RECOGNITION

There are usually two ways to perform target recognition. One is multi-scale window detection means, which in turn extracts fixed shape, different scale windows from the image and their features are sent into classifier to determine whether it is a target; The other is the segmentation means, which segments the image into different areas first, then determines the interest region, extracts the features, then classifier is used to determine whether it is a target.

According to the scene semantics to be identified, the related interested targets are determined by expert system and then are extracted respectively. The existing target recognition classifier, such as support vector machine and neural network, often give an output value O_i and then comparison is performed with fixed threshold t_i and the result is "either-or". It is difficult to select a threshold for comparison for this method is easy to cause the false positive and false negative. In order to enhance the follow-up processing robustness, here to the output of target recognition classifier is fuzzed, namely represent recognition results using membership degree. According to the characteristics of output decision value, the form of membership functions is as follows:

$$A_i(x)=\begin{cases}\min(1,1-e^{(-\frac{o_i-t_i}{T_i})}+T_\mu), & \text{if } o_i-t_i\geq 0\\ T_\mu, & \text{else}\end{cases} \tag{11}$$

where, A_i represents the first i kind target set, x represents target, $A_i(x)$ represents the membership degree, T_i is an constant, T_u is the minimum membership degree.

In order to save calculation time, it is necessary to avoid getting too much targets. For the first i kind targets, we only keep at most N_i targets that membership degree is bigger than T_u, so the target recognition result is a target set:

$$X_i=\{x_{ij}\mid A_i(x_{ij})>T_\mu,\ j=1,2,\ldots,N,\ 0\leq N\leq N_i\}$$

and then attribute extraction is performed.

Attribute extraction refers to extract the location, size and other information of the target. For multi-scale window detector, the center of the detected window is the target position and the length together with the width of the detected window represents the target size information. For segmentation detector, the center of the Minimum Boundary Rectangular (MBR) is the position of the target and the length together with the width of the MBR represents the target size information. These Attributes are extracted to perform the fuzzy reasoning, such as analysis the spatial relations between targets. Each target class has an attribute set B_i, $B_i=\{y_{ij}\mid j=1,2,\ldots n_i\}$.

FUZZY REASONING

Scene semantics is sentenced by fuzzy rule base of expert system. The inputs of fuzzy reasoning machine are membership degrees and attributes of all targets and the output is the membership degrees to the current scene semantics. The form of fuzzy rules is as follows:

$$\begin{gathered}\textit{if}\ (A_1(x_1)\text{ is }C_1)\text{ and }(A_2(x_2)\text{ is }C_1)\text{and}(f_1(x_1,x_2)\text{ is }C_1)\text{ then }D_l\ ;\\ \textit{if}\ (A_1(x_1)\text{ is }C_1)\text{ and }(A_2(x_2)\text{ is }C_2)\text{and}(f_1(x_1,x_2)\text{ is }C_2)\text{ then }D_2\ ;\\ \textit{if}\ (A_1(x_1)\text{ is }C_2)\text{ and }(A_2(x_2)\text{ is }C_1)\text{and}(f_1(x_1,x_2)\text{ is }C_1)\text{ then }D_3\ ;\end{gathered}$$

where, x_i (i = 1, 2, …, n) represents the detected first i kind target, f_k (x_i , x_j) (i, j, k = 1, 2, …n) represents the first k fuzzy relation measure function of x_i and x_j and its scope is [0, 1]; C_k (k = 1, 2, …, n) is fuzzy set that represents the intensity of the relation and its scope is [0, 1]; D_k (k = 1, 2, …, n) is fuzzy set that represents the conclusion about scene semantics and its scope is [0, 1]; For simplified calculation, the membership function of C_k and D_k is triangle function. According to the important degree of knowledge, the fuzzy rules are given different credibility, which helps launch reliable conclusion.

When perform fuzzy reasoning, take out a target and its attributes from each target set X_i as the input of the automata. Here the former part matching of the rule is performed, if none target is detected, namely N = 0, then the related rules will not participate in the fuzzy reasoning. Fuzzy operation adopts Single-point fuzzy method and operation adopts minimum calculation rule, implication operation adopts Mamdani-minimum calculation rule, synthetic operation adopts the maximum-minimum calculation rules and clearness operation adopts center of gravity method (COG).

Suppose there are k target class and at most Ni targets is detected for first target class i, then there are $\prod_{i=1}^{k} N_i$ output values at most, take a maximum of these output values as the ultimate decision value.

EXPERIMENT

Suppose the wanted scene semantics I is "Referee shows a red card in the football match", which is shown in Fig. 2. Through human experience it could be easily known that in this scene: There must be the referee's handheld red card, in addition to a great degree the face or the human body of referee or players will appear; Handheld red card is higher than face and is similar with face in size, etc. According to the experience we can determine the interested target set and establish the fuzzy rules and its credibility. There are 3 interested target sets, i.e., handheld red card A_1, face A_2, stand human body A_3. x_1, x_2, x_3 represents every kind of targets. C_1, C_2, C_3 represents that membership degree is weak, general, strong respectively; D_1 represents it is

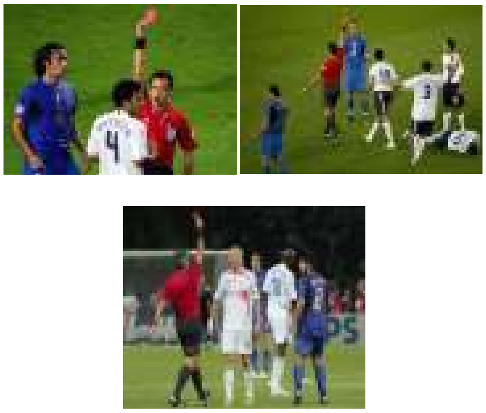

Fig. 2: Referee shows a red card in the football match

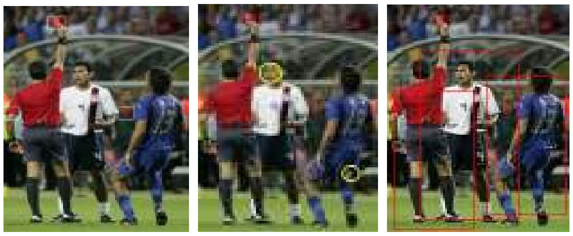

Fig. 3: Target detectors

impossible that scene semantics is *I*, D_2 represents it is possible that scene semantics is *I*, D_3 represents it is extremely possible that scene semantics is *I*.

Three detectors are trained respectively, i.e., handheld red card detector, face detector and human body detector. Handheld red card detector takes the RGB averages of image sub-block as features, face detector takes the Haar wavelet characteristics as features and human body detector adopts the HOG features. Three kinds of detector use multi-scale window measure method and the detection result is shown in Fig. 3. It is obviously that false negative is eliminated by fuzzing the detection results. There is false positive in human body detection and face detection and the detected position of human body is not particularly accurate.

Ninety images are collected as positive samples from internet that the image semantics is "Referee shows a red card in the football match" and negative samples are 1000 images selected from INRIA database, the content of which include landscape, characters, sports, etc. Multilayer filter method is usually adopted for identifying semantics of specific situations and achieved satisfied effect (Lijuan *et al*., 2002; Yin *et al*., 2004). A corresponding decision diagram for semantics recognition is established according to the thought of multilayer filter, which is shown in Fig. 4.

The Precision and Recall are two universal criterions to evaluate the performance of recognition algorithm. Assume the images which actually accord with the semantics is {Relevant}, the images which are sentenced to accord with the semantics by recognition system is {Retrieved} and the images which not only actually accord with the semantics but also are sentenced to accord with the semantics by recognition system is {Relevant}∩{Retrieved}, then Precision is:

$$\text{Precision}=\frac{|\{\text{Relevant}\}\cap\{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|} \tag{12}$$

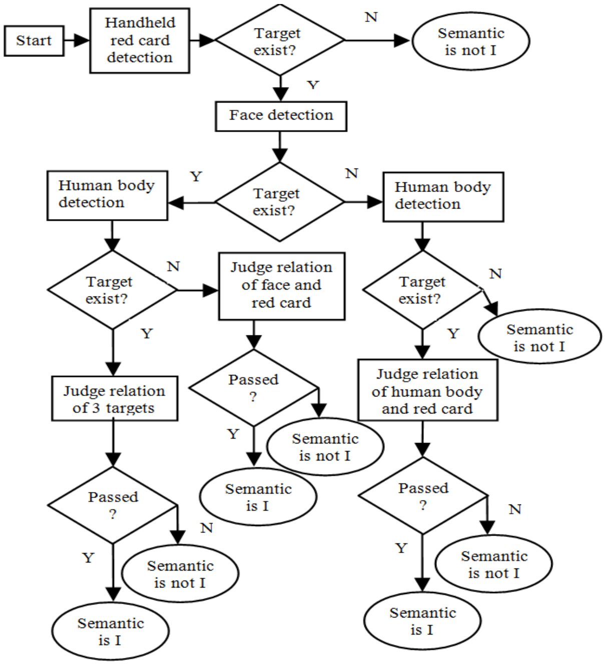

Fig. 4: Decision diagram for semantics recognition

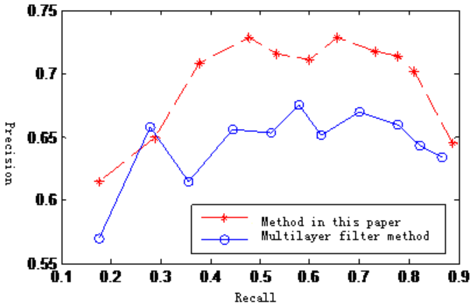

Fig. 5: Comparison of two methods

which indicates the recognition system's ability to reject the negative samples. Meanwhile Recall is:

$$\text{Recall}=\frac{|\{\text{Relevant}\}\cap\{\text{Retrieved}\}|}{|\{\text{Relevant}\}|} \tag{13}$$

which indicates the recognition system's ability to accept the positive samples. Higher Precision and Recall means the better performance of the recognition system.

Experiments of Comparison of the two methods are performed using MATLAB and OpenCV on a computer and Fig. 5 shows the evaluation curve of recognition effect.

It is obviously that the method presented by this study is better than multi-level filter method on the whole. This is because when the threshold in multilayer filter is bigger, false negative happens, which causes a low recall. On the contrary, when the threshold in multilayer filter is smaller, false positive happens, which causes a low precision. The method in this study fuzz the outputs of the target detector first, hence false negative is avoided; Then fuzzy reasoning is performed to reduce the false positive, so better result is obtained.

CONCLUSION

In order to get better image semantics recognition, in this study, a scene semantics recognition system based on fuzzy reasoning is presented. The system contains four parts: image preprocessing, feature extraction, target recognition and fuzzy reasoning machine. Compared with other methods, the outputs of target detectors are fuzzed, the fuzzy relationships between targets are extracted and fuzzy inference is performed using fuzzy automata. The experiment indicates that this method could overcome the problems of false positive and false negative of pattern classifiers and perform relatively more accurate image semantics recognition than other existing methods.

ACKNOWLEDGMENT

This study was supported by "the Fundamental Research Funds for the Central Universities" of North China Electric Power University and State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources.

REFERENCES

Boles, W. and B. Boashah, 1998. A human identification technique using images of the iris and wavelet transform. IEEE T. Signal Proces., 46(4): 1185-1188.

Dalal, N., 2006. Finding people in images and videos. Ph. D. Thesis, Institute National Polytechnique de Grenoble.

Dalal, N. and B. Triggs, 2005. Histograms of oriented gradients for human detection. IEEE International Conference on Computer Vision and Pattern Recognition, pp: 886-893.

Li, F.Z. and K.H. Hu, 2005. Proceeding of non-rigid motion analysis. J. Image Graph., 10(1): 12-13.

Lijuan, D., C. Guoqing, G. Wen and Z. Hongming, 2002. A hierarchical method for nude image filtering. J. Comput. Aid. Design Comput. Graph., 14(5): 404-409.

Tan, T.N., 1995. Texture edge detection by modeling visual cortical channels. Pattern Recogn., 28(9): 1283-1298.

Viola, P., 2001. Rapid target detection using a boosted cascade of simple features. Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, pp: 511-518.

Yin, X.D., D. Tang, J. Deng, *et al*., 2004. Content-based method for nude image filtering. Comput. Automat. Measurement Control, 12(3): 283-286.