International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.8 (2015), pp.79-88
http://dx.doi.org/10.14257/ijsip.2015.8.8.09

Image Cropping by Patches Dissimilarities

Shangbing Gao [*] [1,2], Youdong Zhang [1], Wanli Feng [1] and Dashan Chen [2]

1. *Faculty of Computer Engineering, Huaiyin Institute of Technology, Huai'an, China*
2. *The Key Laboratory for traffic and transportation security of Jiangsu Province, Huai'an, China*

Email: luxiaofen_2002@126.com, zhangyoudong@hyit.edu.cn, fengwanli@sina.com, chendashan@hyit.edu.cn

Abstract

Image cropping is a technique to help people improve their taken photos' quality by discarding unnecessary parts of a photo. A novel method is presented for Image cropping by patches dissimilarities. In this paper, we propose a novel patches dissimilarities algorithm. Firstly, representing the image patches and reducing dimensionality. Then we extract the visual saliency map of these photos based on the patches dissimilarities. Finally, by the saliency map and face priors, we find a cropped region that can be best found. The experimental results demonstrate that our technique is applicable to a wide range of photos and produce more agreeable resulting photos.

Keywords: *Terms—Image Cropping, Reduce Dimensionality, Patches Dissimilarities*

1. Introduction

Aspect ratio of an image is a concept used to describe the ratio of the width of the image to its height. Image rendering or capture devices (*e.g.* widescreen TV, computer LCD monitors) typically adopt a specific aspect ratio from a set of about half a dozen possibilities.

Nevertheless, three common aspect ratios are most frequently adopted: 4:3 (*e.g.* standard TV broadcast, CRT monitors), 3:2 (*e.g.* photographic paper), 16:9 (*e.g.* widescreen TV, some computer LCD monitors), and usually a means of conversion between these formats is required for media adaptation between devices. Three traditional methods for conversion between different aspect ratios are: Pillars, Stretch and Zoom. The Pillars method leaves solid color bars (also called pillars) in each side of the image. The Stretch method stretches images (*e.g.* using interpolation) to fit into the new proportion. Finally, the Zoom method maintains the aspect ratio by zooming in and cropping the image. The above conversion methods often lead to problems. The first two methods may produce images that are distorted or have undesired borders. The third method preserves the image's aspect ratio, but, if no information is known about the image, the image subject might be cropped.

Since this last method is often chosen by photo printing companies for image adaptation, in this work it will also be referred to as automatic zoom and crop method.

Nowadays, people can use several techniques to improve their taken photos' quality. Cropping is one of such tools –people may want to discard the blurred or noisy parts of a photo, or to emphasize the central objects by extracting and cropping out the most important part of a photo. However, because of the growing number of cameras and thus that of photos taken, people may find it time-consuming and dull to manually crop their snapshots. Automatic photo cropping can free people out of such onerous work.

Various automatic cropping techniques have been proposed, which we will have a brief review in Section II. Previous researches mainly focus on cropping out the important

ISSN: 2005-4254 IJSIP

objects in a photo regardless of their aesthetical quality, *e.g.*, [1-3]. That is, these techniques mainly determine their cropped region by including these important objects. More recent work begins to emphasize the agreeability of the resulting photos by introducing certain photographic rules, *e.g.*, [4-7]. One of them is the Rule of Thirds, i.e., dividing a photo into nine equal-size areas by two horizontal lines and two vertical lines, the subjects of the photo should be centered at one of the four intersections of the four lines.

Within the context of Digital Photography, the main goal of this work is to automatically adapt photographs from any given aspect ratio to a desired proportion while preserving, as long as it is possible, the original photographic composition. Since the results of this work have strong association with photographic development (or printing), large distortions when adjusting for a new aspect ratio are undesired, since they compromise the photographic contents.

This work aims at overcoming the above problems by means of a zooming and selective cropping method. The idea is to consider high and low level information regarding the scene to perform the crop. Therefore, the main goal of the approach is to crop a region to a desired proportion by (1) maximizing the cropped area and (2) avoiding subject chopping. At this stage, our work is restricted to portraits3, since portrait photographics a popular branch of photography.

In order to avoid cropping the subject, patches dissimilarities based the saliency map is used to identify regions within the image that have information that is considered most relevant. For "Face" images, semantic priors such as face priors are very important.

Our work here, photo cropping based on composition, is the work that focuses more on the aesthetical composition of photographs. However, unlike other work that applies specific photographic rules which may not be adaptive to a wide range of photos, we propose to mine the composition data of a large number of high-quality photos and learn to crop automatically.

The paper is organized as following. We will first introduce previous automatic cropping techniques in Section II. In Section III, we explain our proposed algorithm in detail. Finally, we show results of our experiments in Section IV and conclude with further discussion in Section V.

2. Related Works

2.1. Image Cropping

Among previous cropping techniques, some do not pay sufficient attention to aesthetical values of a photo, which merely crop out the major objects based on a saliency map. Ciocca *et al.* [1] used a CART classifier to classify a set of photos into three categories (landscape, close-up, and other). Different modifications are applied to different categories, with the main idea that the cropping result should include the focused elements. Stentiford [2] also cropped the photo mainly based on a saliency map. Santella *et al.* [3] employed eye tracking to help determine the content area for cropping. Such work may be efficient for object-oriented cropping, but they ignore the aesthetical values of photos and may not be applicable to professionally photographic cropping. In a subjective evaluation, the method showed to be almost as effective as manual cropping. Although the results are very good, the method is not fully automatic, requiring user interaction for each image.

In some other work, aesthetical evaluation is emphasized. The most frequently applied standards include color, lighting and composition. To improve photos' composition, Luo *et al.* [4] and Bhattacharya *et al.* [5] applied the Rule of Thirds. In addition to the Rule of Thirds, Liu *et al.* [6] applied the diagonal dominance, visual balance and sizes of salient regions for equally evaluation. Some other features, including spatial distribution of

edges, color distribution, hue count, blur, contrast and brightness (Ke *et al.* [7]), were also used.

Taking these aesthetical evaluations into account, some techniques were proposed. Nishiyama *et al.* [8] trained a SVM to label subject regions of a photo as of high or low quality. Fitting the quality values to the Sigmoid function, they obtained a final quality score by combing the posterior possibilities, and cropped the region with the highest score. Zhang *et al.* [9] proposed three models – composition sub model, conservative sub model and penalty sub model - and combines them linearly to an object function to determine the cropped region.

2.2. Saliency Detection

During last two decades, visual saliency detection and saliency map generation aiming to find out what attracts human's attention got broad interesting in computer vision, especially for object detection or recognition from different scenes. In the following, we introduce some models which are used in our experiments for comparison.

Itti *et al.* [10] introduced a saliency model which was biologically inspired. Specifically, they proposed the use of a set of feature maps from three channels as intensity, color, and orientation. The normalized feature maps from each channel were then linearly combined to generate the overall saliency map. Even though this model has been shown to be successful in predicting human fixations, it is somewhat ad-hoc in that there is no objective function to be optimized and many parameters must be tuned by hand. With the proliferation of eye-tracking data, a number of researchers have recently attempted to address the question of what attracts human visual attention by being more mathematically and statistically precise.

Bruce and Tsotsos [11] modeled bottom-up saliency as the maximum information sampled from an image. More specifically, saliency is computed as Shannon's self-information $-\log p(f)$, where f is a local visual feature vector (i.e., derived from independent component analysis (ICA) performed on a large sample of small RGB patches in the image.) The probability density function is estimated based on a Gaussian kernel density estimate in a neural circuit.

Oliva and Torralba [12] proposed a Bayesian framework for the task of visual search (i.e., whether a target is present or not.) They modeled bottom-up saliency as $1/p(f \mid f_G)$ where f_G represents a global feature that summarizes the appearance of the scene and approximated this conditional probability density function by fitting to multivariate exponential distribution. Zhang et al. [13] also proposed saliency detection using natural statistics (SUN) based on a similar Bayesian framework to estimate the probability of a target at every location. They also claimed that their saliency measure emerges from the use of Shannon's self-information under certain assumptions. They used ICA features as similarly done in [14], but their method differed from [15] in that natural image statistics were applied to determine the density function of ICA features.

Most of the methods [15, 16] based on Gabor or DoG filter responses require many design parameters such as the number of filters, type of filters, choice of the nonlinearities, and a proper normalization scheme. These methods tend to emphasize textured areas as being salient regardless of their context. In order to deal with these problems, [17] adopted non-linear features that model complex cells or neurons in higher levels of the visual system. Kienzle *et al.* [18] further proposed to learn a visual saliency model directly from human eye tracking data using a support vector machine (SVM).

Different from traditional image statistical models, a spectral residual (SR) approach based on the Fourier transform was recently proposed by Hou and Zhang [19]. Spectral residual does not rely on parameters and detects saliency rapidly. In this approach, the difference between the log spectrum of an image and its smoothed version is the spectral

residual of the image. However, Guo and Zhang [20] claimed that what plays an important role for saliency detection is not SR, but the image's phase spectrum.

3. Image cropping via patches dissimilarity

In this section, we will state the framework of our image cropping method in details. The steps of our algorithm are fourfold: representing the image patches, reducing dimensionality, evaluating the patches dissimilarities, image cropping. We will describe the details step by step in the following subsections.

3.1. Representing the Image Patches

The first step of our algorithm is dividing each original input image into small image patches for gathering local information. For simplicity, we take image patches from the original images without overlapping. Given an image I with dimension $H \times W$, non-overlapping patches with the size of $n \times n$ pixels are drawn from it. The total number of patches is $L = \lfloor H/n \rfloor \cdot \lfloor W/n \rfloor$. Denote the patch as $p_i, i = 1, 2, ..., L$. Then each patch is represented as a column vector x_i of pixel values. The length of the vector is $3n^2$ since the color space has three components. Finally, we get a sample matrix $X = [x_1, x_2, ..., x_L]$, L is the total number of patches as stated above.

3.2. Reducing Dimensionality

To effectively describe patches in a relatively low dimensional space, we use an equivalent method to PCA to reduce data dimension. Each column in the matrix X subtracts the average along the columns. Then, we calculate the co-similarity matrix $A = (X^T X)/L^2$, so the size of the matrix A is $L \times L$ .The eigenvalues and eigenvectors are calculated based on the matrix A selected with their eigenvector $U = [u_1, u_2, ..., u_d]^T$ according to the biggest d eigenvalues, where u_i is an eigenvector. The size of the matrix U is $d \times L$.

3.3. Patches Dissimilarity

In the proposed algorithm, the saliency value of each image patch is determined by two factors: one is the dissimilarities between image patches in a reduced dimensional space; the other is the spatial distance between an image patch and all other patches.

With the increasing of the spatial distance between two patches, the influence of the dissimilarity between them was decreasing. Therefore, the dissimilarities were inversely weighted by their corresponding spatial distances. Furthermore, the distance of each patch from the center of the image is involved in the evaluation of the saliency because of the central bias as stated in [3]. With the increasing of the distance between a patch and the center, the saliency of the patch should be appropriately depreciated.

By integrating the elements of dissimilarity, spatial distance and central bias, the saliency of the patch p_i is defined as follows:

$$Saliency(i) = w_2(i) \cdot \sum_{j=1}^{L} \left\{ w_1(i, j) \cdot dist_{color}(p_i, p_j) \right\} \quad (1)$$

where $w_1(i, j)$ is defined as

$$w_1(i, j) = \frac{1}{1 + Dist(p_i, p_j)} \quad (2)$$

Let $dist(p_i, p_j)$ be the Euclidean distance between the positions of patches p_i and p_j, which is represented by the two centers of patches p_i and p_j in the original image, normalized by the larger image dimension.

Thus, a patch p_i is considered salient if the appearance of the patch p_i is distinctive with respect to all other image patches. Specifically, let $dist_{color}(p_i, p_j)$ be the distance between the patches p_i and p_j in the reduced dimensional space. Patch p_i is considered salient when $dist_{color}(p_i, p_j)$ is high for $\forall j$.

$$dist_{color}(p_i, p_j) = \sum_{n=1}^{d} \left| u_{ni} - u_{nj} \right| \tag{3}$$

$w_2(i, j)$ is the second weighting mechanism we proposed according to the average saliency map from human eye fixations indicating a bias to the center of image.

$$w_2(i, j) = 1 - DistToCenter(p_i)/D \tag{4}$$

where $DistToCenter(p_i)$ is the spatial distance between two centers of patch i p and the patch at the center of the original image, and $D = \max_j \{DistToCenter(p_i)\}$ is a normalization factor.

To evaluate a patch's uniqueness, we can compute the dissimilarity between the patch and all of other patches and take the sum of these dissimilarities as the saliency value of the related patch. In practice, there is no need to incorporate its dissimilarity to all other image patches. It suffices to consider the K most similar patches that if the most similar patches are highly different from p_i, then clearly all image patches are highly different from p_i. Hence, for every patch p_i, we search for the K most similar patches $\{q_i\}, i = 1, 2, ..., K$ in the image, according to Equation (2). Under this definition, our method, which measures the saliency from the perspective of global information and local information, is different from existing methods which only consider local contrast or global contrast. A patch p_i is salient when $dissimilarity(p_i, q_k)$ is high for $\forall k \in [1, ..., K]$. The saliency value of patch p_i can be expressed as follows (we choose $K = 100$ in our experiments):

$$S_i = 1 - \exp\{-\frac{1}{K} \sum_{k=1}^{K} dissimilarity(p_i, q_k)\} \tag{5}$$

As described previously, the patch size would influence the calculation of saliency map. With a smaller patch size, the saliency map will become distinguishable, as shown in Figure 1 where the saliency map with the smallest image patch size (shown in Figure 1 (d)) is more distinguishable than the other two with larger patch size (shown in Figure 1 (b) and (c)). Of course, to obtain more accurate saliency map, we hope to divide image into smaller image patches. But, in this situation, the computational complexity will increase. The computational complexity of our algorithm includes twofold: first is the computational complexity on preprocessing, such as dividing original images into patches and PCA; the other time consuming cost is computing dissimilarities between patches. Given an input image with size of $H \times W$ (where H is the height and W is the width) and the patch size of $n \times n$, the computational complexity of our algorithm is $(L^3 + L^2)$, in which L^3 and L^2 corresponding to the computational cost of preprocessing and dissimilarity calculation respectively, where $L = \lfloor H / n \rfloor \cdot \lfloor W / n \rfloor$. Therefore, with the smaller patch size, the computational complexity will increase.

The approach begins by generating a saliency map $Saliency(x)$ and then searching for the window which possesses the highest average pixel attention score according to equation (5). Fig. 1 shows an image together with its saliency map in which the brighter pixels indicate higher saliency.

(a) Original image (b) Saliency map

Figure 1. The Corresponding Saliency Map

3.4. Image Cropping

People pay more attention to certain semantic objects such as faces even without specific purposes.

Therefore, similar as in [21], we perform face detection on the images. The regions near the detected faces are assigned higher priors $p_f(x) = \exp(-d(x, f_c)/\sigma_2^2)$, where f_c is the center of the face.

1) For "Face" images, we use the abovementioned method to get the saliency map;
2) For "Other" images, saliency map is generated based on patches dissimilarites model. Visual attention facilitates the processing of the portion of the input associated with the relevant information, suppressing the remaining information.
3) The saliency map is automatically binarized in order to identify saliency regions. The regions with areas smaller than a threshold which are a function of the area of the larger region are discarded.
4) 4)A single relevant region is obtained, considering the bounding box that includes all the saliency regions previously identified.
5) The image is then cropped and adapted with respect to this region.

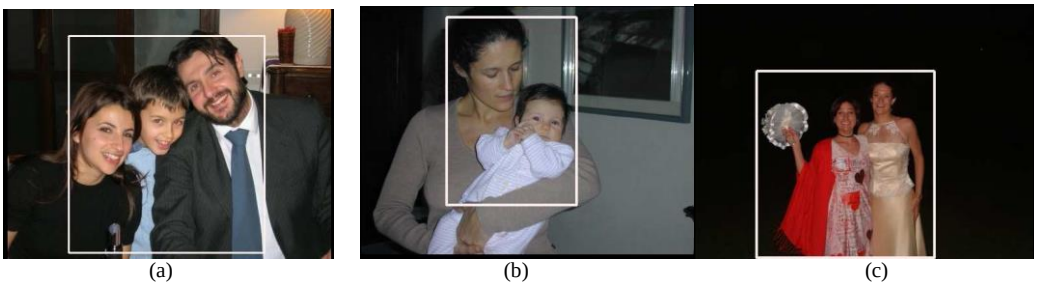
(a) Original image (b) The corresponding cropped version

Figure 2. Original Image and Cropped Version

4. Experimental Results

4.1. Image Cropping Results

The window W may be any shape or size, but for many image forming devices a fixed aspect ratio may be required and the method is illustrated here. Of course, faces are very important. Thus, faces are salient. Figure 3 shows relevant regions selected within some of the "Other" images containing faces. Figure 4 shows a series of images together with their cropped versions.



(a) (b) (c)

Figure 3. Relevant Regions Selected within Some of the "Other" Images Containing Faces

(a) original image (b) cropping result (c) original image (d) cropping result

Figure 4. User Evaluation

4.2. User Studies

In the experiment, we use the dataset of [6] with a total of 93 photos for testing, and it takes about 10 seconds to process each photo. Some of the results are reflected in Figure 2 - 3. We randomly chose 11 photos and compare our results with that of Santella [3] for user studies. We designed a questionare. For each question, we include the original photo as reference, and list our resulting photo and [3] resulting photo as options. Users are asked to choose one of the two that is of better composition. The order of display of the two options is at random, and users have no idea of which photo is our result. We post the questionare online, and receive 91 replies.

We show some of the survey results in detail from Figure 5 to Figure 6. We display the original testing photos in column (a), [3]'s results in column (b), our results in column (c),

and the survey results in column (d). As reflected from Figure 5 and Figure 6, our results are much more favored than [3]'s results, because our method focuses more on composition of photos as a whole.

(a) original image (b) Santella's [3] result (c) our result; (d) Percentage of subjects who choose (b) or (c) as the better result

Figure 5. User Studies

(a) original image (b) Santella's[3] result (c) our result; (d) Percentage of subjects who choose (b) or (c) as the better result

Figure 6. User Studies

4.3. Comparison with Related Work

(a) original image (b) Santella's[3] result (c) our result

Figure 7. The Comparison of the Two Algorithms

(a) original image (b) Santella's [3] result (c) our result

Figure 8. The Comparison of the Two Algorithms

We compare our results with some other work in this subsection. We show that our cropped results focus on aesthetical values as Figure 7 shows. Unrelated objects are

discarded. And if the original photos are good enough, we can leave them alone as Fig. 8 shows. Also, as our results reflected, our technique is adaptive to a wide range of photos, including natural scenes, animals, human profile, etc.

5. Conclusions

This paper has described the application of an attention measure to the automatic cropping of images. The method may be applied to cropping single images or guiding a series of zoom operations. We show that our technique generates resulting photos with better composition. Also, our technique can be applied to different types of photos. In our future work, we aim to improve our scene classification, mine composition from more photos, and accelerate our cropped region searching method.

References

[1] [1] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini. Self-Adaptive Image Cropping for Small Displays.IEEE Transactions on Consumer Electronics, 2007 , 53(4): 1622–1627.
[2] [2] F. Stentiford. Attention Based Auto Image Cropping.In ICVS Workshop on Computational Attention & Applications, 2007.
[3] [3] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. F. Cohen.Gaze-based interaction for semi-automatic photo cropping.In ICVS Workshop on Computational Attention & Applications, 2007. [4] Y. Luo and X. Tang.Photo and video quality evaluation: Focusing on the subject. In Proceedings of the 10th European Conference on Computer Vision: Part III, ser. ECCV '08, 2008,386–399.
[4] [5] S. Bhattacahrya, R. Sukthankar, and M. Shah.A framework for photoquality assessment and enhancement based on visual aesthetics. In Proceedings of the international conference on Multimedia, ser. MM'10. ACM, 2010, 271–280.
[5] [6] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. Computer Graphic Forum (Proceedings of Eurographics), 2010, 29(2): 469–478.
[6] [7] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. in Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2006,1:419–426.
[7] [8] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensation-based photo cropping. in Proceedings of the 17th International Conference on Multimedia 2009, Vancouver, British Columbia, Canada, 2009. ACM, pp. 669–672.
[8] [9] M. Zhang, L. Zhang, Y. Sun, L. Feng, and W. Ma.Auto Cropping for Digital Photographs. In IEEE International Conference on Multimedia and Expo, 2005.
[9] [10] L. Itti, C. Koch and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Analysis and Machine Intelligence, vol.20, no.11, pages 1254-1259, 1998.
[10] [11] N. Bruce and J. Tsotsos. Saliency based on information maximization. In Advances in Neural Information Processing Systems, pages 155-162, 2006.
[11] [12] A.Oliva, A. Torralba,M. Castelhano, and J. Henderson. Top-down control of visual attention in object detection. In Proceedings of International Conference on Image Processing, pages 253-256, 2003.
[12] [13] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision, 8(7):32, pages 1-20, 2008.
[13] [14] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In Advances in Neural Information Processing Systems, pages 481-488, 2004
[14] [15] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to photograph. in Proceedings of the international conference on Multimedia, ser. MM'10. ACM, 2010, 291–300.
[15] [16] F.-F. Li and P. Perona.A Bayesian Hierarchical Model for Learning Natural Scene Categories. in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, 2:24–531.
[16] [17] J. Harel, C. Koch, and P. Perona.Graph-based visual saliency. in Advances in Neural Information Processing Systems,2007 , 19:545–552.
[17] [18] W. Kienzle, F. Wichmann, B. Scholkopf, and M. Franz. A nonparametric approach to bottom-up visual saliency. In Advances in Neural Information Processing Systems, pages 689-696, 2007.
[18] [19] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1-8, 2008.
[19] [20] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1-8, 2008.
[20] [21] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In ICCV, 2009.

Acknowledgements

This work is sponsored by the National Natural Science Foundation of China (NSFC) #61402192, JiangSu Qing Lan Project, six talent peaks project in Jiangsu Province #2013DZXX023, Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 14KJB520006), Jiangsu 333 Project, the Science & Technology Fund of Huai'an under the Grant No.HAG2013059, HANZ2014006, the open fund for the Key Laboratory for traffic and transportation security of Jiangsu Province (TTS2015-05), the open fund of Jiangsu Provincial Key Laboratory for advanced manufacturing technology (HGAMTL-1401), the open fund of Jiangsu Provincial key laboratory for interventional medical devices (JR1405).

Authors

Shangbing Gao, he was born in 1981, received the BS degree in mathematics from the Northwestern Polytechnical University in 2003. He received the MS degree in applied mathematics from the Nanjing University of Information and Science and Technology in 2006. He received the Ph.D. degree with School of Computer Science and Technology, Nanjing University of Science and Technology (NUST). Since 2014 he has been an associate professor at Huaiyin Institute of Technology, China. Her current research interests include pattern recognition and computer vision.

Youdong Zhang, he was born in 1967, received the BS degree in Physics from the Nanjing Normal University in 1989. He received the MS degree in computer application from Nanjing Aeronautics and Astronautics University in 2000. He received the Ph.D. degree in Digital engineering and information security from Nanjing Aeronautics and Astronautics University. He is a professor at Huaiyin Institute of Technology, China. Her current research interests include digital Forensic and computer vision.

Wanli Feng, he was born in 1973, received his B.S. degree in Software Engineering from Tsinghua University, Beijing, China in 2003, his M.S. degree in computer science and technology from Southeast University, Nanjing, China in 2010. He is an associate professor at Huaiyin institute of technology, Huai'an, China. His current research interests include image processing and software.

DaShan Chen, he aws born in October 1983, received the B.S. and M.S. degrees from Chang'an University. He received the Ph.D. degrees from Tongji University, Shanghai, China, in 2012. He is currently a lecturer in the school transportation engineering, Huaiyin institute of technology. His research interests include traffic safety an

