A VOTE OF CONFIDENCE BASED INTEREST POINT DETECTOR

Zhenwei Miao and Xudong Jiang

School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore 639798
{mi0001ei, exdjiang}@ntu.edu.sg

ABSTRACT

In this paper, a vote of confidence (VC) based detector is proposed to detect bright and dark regions from images. Whether a local region is bright or dark is voted by all the pixels in this region. Compared to the contrast based detectors, such as the popular SIFT detector, the VC detector is invariant to illumination change and robust to abrupt variations. Experiments are conducted on benchmark databases to verify the superior performance of the VC detector in terms of the repeatability and matching score. The proposed detector is also evaluated in the application of face recognition.

Index Terms— Interest point detection, vote of confidence, image matching, repeatability, face recognition.

1. INTRODUCTION

Interest points detection is an important research topic in image processing, analysis and recognition [1–7]. It provides a way to represent images with sparse local patches, and has been proven to be well suitable to deal with the challenges of clutter, occlusion and variations of viewing condition [8]. Various detectors have been developed in the last decades [9–26], most of which are designed directly based on image contrast. The representative ones include the Harris [9, 10], Hessian [13], Harris Laplace/affine [11], Hessian Laplace/affine [11], SIFT [14] and SURF [15] detectors. They detect interest points from the responses derived from the first or second order derivative of image intensity. Such detectors prefer the local structures with high contrast. Low contrast structures will not be easily detected even if they are stable under difference variations. Moreover, the first or second order derivative amplifies the image noise. This causes these detectors sensitive to noise. The ROLG detector [19,27] uses the rank order filter instead of the linear filter to reduce the influence of noise and the nearby structures. However, it still prefers the structures which have high contrast.

Detectors which are not directly designed from image contrast are also developed. The MSER [17,20], MSCR [21], PCBR [22] and BPLR [23] detectors are based on the image segmentation algorithms. As image segmentation is still a challenging task, the performance of these detectors becomes poor under image blurring in which the boundaries of structures turn to unclear [3]. The statistical properties of local regions are employed by the SUSAN [12], FAST [24], and salient region [25, 26] detectors. Both the SUSAN and FAST detectors use the similarity between the nucleus (central pixel) and its surrounding pixels to generate the corner map. As the corner maps of the both detectors are generated from local regions with fixed size, these two detectors are not scale invariant. The number of interest points detected by the salient region detector is small due to the greedy cluster method used to group the nearby interest points [3].

Instead of computing the corner map directly from the image contrast (e.g. SIFT [14]), or using the image segmentation algorithm (e.g. MSER [17]) or based on the local statistical property (e.g. salient region detector [25]), we proposed a vote of confidence (VC) based detector in this paper. Each local region is separated into a concentric ring and circle. Mutual voting is conducted by these two parts. Interesting points are detected from the map of the voting confidence score. The proposed VC detector is robust to illumination changes and effective to cluttered structures.

2. VOTE OF CONFIDENCE BASED DETECTOR

Instead of requiring all the pixels in the local regions brighter or darker than their surrounding as done by the MSER detector [17], the brightness/darkness of a local region is measured by a linear combination of the two levels of confidence defined as follows.

Confidence level 1: the normalized number of pixels in a region that are brighter/darker than the majority of the pixels in its surrounding region.

Confidence level 2: the normalized number of pixels in the surrounding region that are brighter/darker than the majority of the pixels in its surrounded center region.

By using these two confidence levels, the measurement of the brightness/darkness has a larger toleration of the illumination variation and abrupt structures than using only one. In order to simplify the problem of local region detection, we restrict the local region to a circle image patch, which is also adopted by many detectors [12, 14, 18, 25]. Each local region

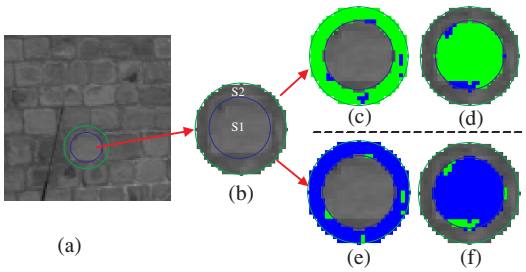

Fig. 1. (a) input image. (b) an enlarged image patch. (c)&(d) voting for brightness: (c) voting results by the pixels in the surrounding part and (d) voting results by the pixels in the inner part. (e)&(f) voting for darkness: (e) voting results by the pixels in the surrounding part and (f) voting results by the pixels in the inner part. For each voting pixel in (c) to (f), green color represents voting by 1 while blue color means voting by 0. Best viewed in color.

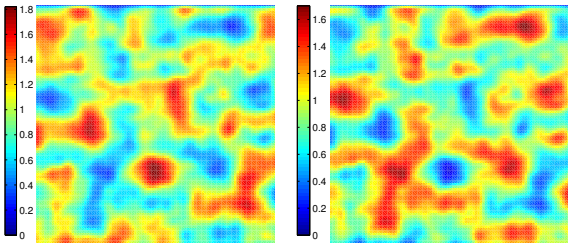

Fig. 2. Voting maps of (a) bright regions and (b) dark regions. Best viewed in color.

(for example, shown in Fig. 1(b)) is separated into two parts: inner circle disk S_1 and its surrounding ring S_2. Confidence levels 1&2 are generated from these two parts.

The vote of confidence (VC) is proposed in Section 2.1 to measure the degree of brightness and darkness. Algorithm to remove the unstable points on ridge is presented in Section 2.2. The VC detector in multiple scales is given in Section 2.3.

2.1. Voting Algorithm

Two quantitative measurements, named VC for brightness and darkness (VCB and VCD), are proposed to measure the degree of brightness and darkness. As the VCB and VCD follow the analogous rules, we take the VCB as an example to derive the voting algorithm.

A bright image patch should be bright in the central region and dark in its surrounding. Therefore, for each image patch, such as Fig. 1(b), the VCB is determined by two parts: the VC that the inner circle S_1 is bright and the VC that its surrounding ring S_2 is dark. The confidence of brightness for S_1 is voted by the pixels in S_2, and similarly the confidence of darkness for S_2 is voted by the pixels in S_1. The following question is that how a pixel votes its counterpart region, for example, how a pixel I_i in S_2 votes for the brightness of S_1? Obviously, if all the pixels in S_1 are brighter than I_i, I_i should vote 1 for the brightness of S_1. However, this makes it sensitive to impulsive noise and abrupt structures. In order to alleviate this problem, in this paper the voting rule is set as follow.

Voting rule: If a pixel I_i is brighter/darker than more than half pixels in the counterpart region S_j, it votes 1 for the darkness/brightness of S_j, otherwise, it votes 0. Let the median for S_j be ϕ_j. The voting rule for brightness is

$$vb(I_i, S_j) = \begin{cases} 1, & \text{if } I_i < \phi_j \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

and that for the darkness is

$$vd(I_i, S_j) = \begin{cases} 1, & \text{if } I_i > \phi_j \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

One example of the voting for the bright region is shown in Fig. 1(c) and (d). In Fig. 1(c) and (d), green color pixels represent voting by 1 while blue color pixels mean voting by 0. Fig. 1(c) depicts the voting results of the pixels in S_2 for the brightness of S_1. As S_1 is bright, the majority of the pixels in S_2 vote 1 for the brightness of S_1. Similarly, in Fig. 1(d) the majority of the pixels in S_1 vote 1 for the darkness of S_2.

The VCB is a linear combination of the normalized voting results for the brightness of the inner circle (Confidence level 2) and that for the darkness of the surrounding ring (Confidence level 1). The VCB at location t is

$$VCB(t) = \sum_{i \in S_2} \frac{vb(I_i, S_1)}{s_2} + \sum_{i \in S_1} \frac{vd(I_i, S_2)}{s_1} \tag{3}$$

where s_1 and s_2 are the area size of S_1 and S_2, respectively. The response of the VCB (named VCB map) for Fig. 1(a) is shown in Fig. 2(a). It is seen that the bright regions have high response while the dark regions have low response.

Similarly, the VCD at location t is defined as

$$VCD(t) = \sum_{i \in S_2} \frac{vd(I_i, S_1)}{s_2} + \sum_{i \in S_1} \frac{vb(I_i, S_2)}{s_1}. \tag{4}$$

The VCD map for the image in Fig. 1(a) is shown in Fig. 2(b). It enhances the dark regions and suppresses the bright regions.

2.2. Ridge Suppression

By detecting the local peaks, interest points are extracted from the VC maps (VCB and VCD maps). If the scale of the image patch matches with the width of the ridge, the VC on this ridge may be larger than 1. In this case, slight vibration may cause false detection of interest points on the ridge. Such kind of unstable interest points need be removed. Although the peaks of the VC response on ridge have large amplitude, the difference between the peak value $R(t)$ and the maximum

value in the corresponding surrounding region S_2 is small. Hence, we employ the ratio

$$\lambda = (R(t) - max\{R(i)|i \in S_2\})/max\{R(i)|i \in S_2\} \quad (5)$$

to remove the unstable interest points on the ridge. If λ is small, it means the peak is very similar to its nearby region. Such interest point candidate is most likely on the ridge. We remove such candidates if $\lambda < 0.05$, which is chosen experimentally.

2.3. VC Detector in Multiple Scales

Interest point detection in multiple scales is an important issue in vision applications. By changing the radius of the local image patches S_1 and S_2, the VC detector achieves the multi-scale detection of local structures. Similar to other detectors, the structures detected by the VC detector appear in a wide range of scales. Sometimes, no sharp maximum is generated along the scale dimension. In this case, the local extremum along the scale dimension is sensitive to noise, and it is unreliable in determining the scale size. From another perspective, as the local structure appears in multi-scales, the result should be more reliable if we use all such scales instead of only one to make a decision. In the following, a grouping method is proposed to cluster the connected interest points.

Let an interest point at location (x_i, y_i) and scale s_i be $\mathcal{P}_i = (x_i, y_i, s_i, R_i, F_i)$ where R_i is the VC response and $F_i \in \{Bright, Dark\}$ is the flag of bright or dark region. The connection of two interest points $\mathcal{P}_i$ and $\mathcal{P}_j$ is defined as: 1) they are both bright or both dark regions, and 2) they are close along both the spatial and scale dimensions. In this work, the distance in the spatial dimension is restricted as $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} < 0.3s_i$, and in the scale dimension s_i and s_j should be the immediately neighbor or only one discrete scale exists between them. All the connected interest points are clustered into one group. Assume a group $\mathbf{G}_p$ contains N interest points as $\mathbf{G}_p = \{\mathcal{P}_{p1}, \mathcal{P}_{p2}, ..., \mathcal{P}_{pN}\}$. A representative interest point $\mathcal{P}_p = (x_p, y_p, s_p, R_p, F_p)$ for this group is generated by setting the location $x_p = \sum_{i=1}^{N} x_i/N$, $y_p = \sum_{i=1}^{N} y_i/N$, the scale $s_p = \sum_{i=1}^{N} R_i^2 s_i / \sum_{i=1}^{N} R_i^2$ and the response $R_p = max\{R_1, R_2, ..., R_N\}$. By weighting the scale with R_i^2, the scales with large VC responses have high influence on determining the scale of the corresponding group.

The proposed algorithm for the VC detector is summarized as follow:

1: Generate the VC response on multi-scales.
2: Detect the local maximums on both the scale and spatial dimensions with some toleration.
3: Remove the points on ridges.
4: Group the interest points which are corresponding to the same structure.

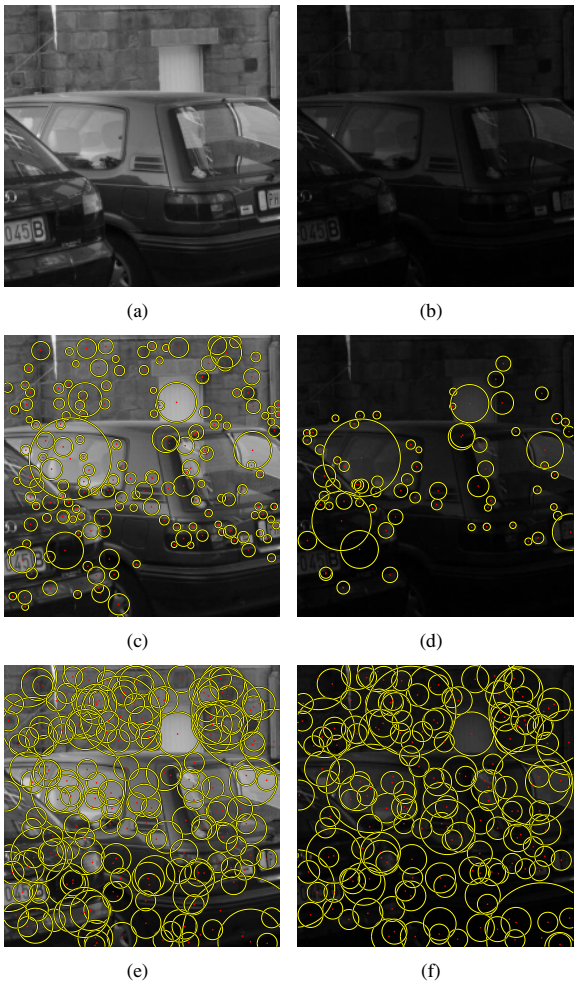

Fig. 3. (a)&(b) input images. (c)&(d) interest points detected by SIFT detector. (e)&(f) interest points detected by VC detector. Best viewed in color.

3. EXPERIMENTS

3.1. Visual Inspection

The images from the Oxford database [2] 'leuven' data set are used to depict some visual results of the VC detector under illumination changes. The setting of different detectors is the same as that given in Section 3.2. It is seen that the number of the interest points detected by the SIFT detector (shown in Fig. 3(c) and (d)) decreases with the scenes darkening. In contrast, the influence of the illumination change has little affect on the VC detector. Most of the structures detected by the VC detector are repeated in these two scenes (shown in Fig. 3(e) and (f)). Besides, although some structures in this scene cluster with each other, the VC detector can still separate them and detect them out.

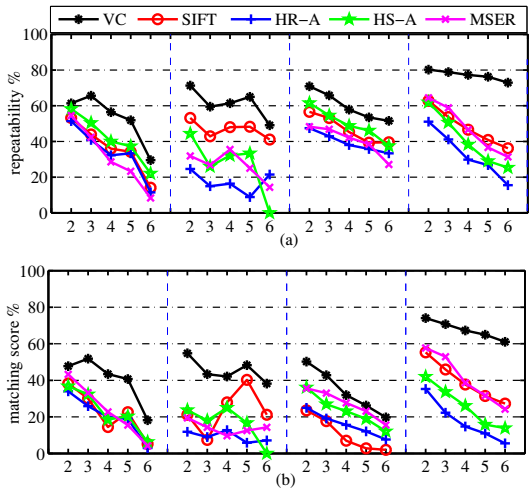

Fig. 4. (a) repeatability and (b) matching score on the Oxford database. In each column, horizontal axis represents the image index in the corresponding data set. From left to right of (a) and (b) are the results on the scale change for structured sequence, the scale change for textured sequence, the blurring for textured sequence and the illumination change sequence, respectively. Best viewed in color.

3.2. Repeatability and Discrimination Tests

The aim of this experiment is to evaluate the detectors under different variations based on the protocol in [2]. Detectors are compared by repeatability and matching score. Two detected regions are repeated if their overlap is above a certain threshold (it is set to be 60% as suggested in [2]). The repeatability/matching score is the ratio between the number of repeated/matched points and the larger number of detected points in the same scenes of each image pair. The test data sets are chosen from the standard publicly available database in [2] .

Similar to that done in [18], interest points are detected on the half-sampled images. For the VC detector, interest points are detected on 5 octaves by half-sampling the previous octave. In each octave, local extrema are detected on 6 scales: $\{\sigma_n\}_{n=1,2,...,6} = \{3, 4, ..., 8\}$. The threshold to remove the low VC points is set to be 1.5. Four detectors, the MSER [17], Harris-affine (HR-A) [11], Hessian-affine (HS-A) [11] and SIFT [14] detectors are compared with the VC detector. The default parameters of these four detectors supplied by authors are employed here. SIFT descriptor [14] is used to describe interest points for all detectors included here. Experimental results are shown in Fig. 4. The VC detector outperforms other 4 detectors in all cases.

3.3. Application to Face Recognition

Face recognition is an active research topic [28–31] and some work has been done to apply SIFT descriptor in face recognition [32, 33]. In this part, the VC detector is compared with the MSER [17], HR-S [11], HS-A [11], and SIFT detectors [14]. As the default setting produces too few interest points for the face recognition for all detectors, the contrast threshold is set to be zero for all detectors in this experiment. For the VC detector, the threshold to remove the low VC points is set to be 1. For the MSER detector, the minimum size of output region is set to be 1/4 of its default setting to make it workable on all face databases. All the detected interest points are described by the SIFT descriptor. The matching algorithm is the one given in [14].

Table 1. Recognition Rate on AR, ORL, GT, and FERET Databases.

	AR	ORL	GT	FERET
VC	96.8%	94.5%	91.1%	96.7%
SIFT	94.3%	90.0%	84.0%	89.9%
HS-A	88.6%	80.0%	74.0%	85.3%
HR-A	74.5%	66.5%	47.4%	49.7%
MSER	92.7%	91.0%	81.1%	89.3%

The AR [34], ORL [35], GT [36] and FERET [37] databases are used to evaluate these detectors. For the AR database, gray images are normalized into the size of 60×85. 75 subjects with 14 nonoccluded images per person are selected. The first 7 images of all subjects are chosen as gallery set, and the remaining 7 images as probe set. Images in the ORL database are normalized into the size of 50×57. The first 5 images of all 40 subjects are chosen as gallery set, and the remaining 5 images as probe set. Gray images in the GT database are normalized into the size of 60×80. The first 8 images of all 50 subjects are chosen as gallery set, and the remaining 7 images as probe set. For the FERET database, images are cropped into the size of 60×80. 1194 subjects with 2 images per person are selected. The first 1 image of all subjects is chosen as gallery set, and the remaining 1 image as probe set. Table 1 shows the recognition rate on the four databases. It is seen that the VC detector significantly outperforms the other 4 detectors over the four databases.

4. CONCLUSIONS

A vote of confidence based detector is presented in this paper to detect bright and dark regions from images. Voting rules are proposed to tolerate the impulsive noise and abrupt structures. Grouping algorithm is designed to determine the scales of local structures. Compared to the LoG filter, the VC response is independent of image contrast and robust to cluttered surrounding. Experimental results demonstrate that the VC detector has better performance in dealing with scale, blurring and illumination changes compared to other 4 detectors in terms of repeatability and matching score. Its superiority is further verified on the experiments of face recognition.

5. REFERENCES

[1] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *Int. J. Computer Vision*, vol. 37, no. 2, pp. 151–172, 2000.

[2] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. van Gool, "A comparison of affine region detectors," *Int. J. Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[3] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Fundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[4] M. Brown and D. G. Lowe, "Recognising panoramas," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 1218 –1225.

[5] R. Fergus, F. F. Li, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *Proc. Int. Conf. Computer Vision*, 2005, pp. 1816 –1823.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Computer Vision*, 1999, vol. 2, pp. 1150–1157 vol.2.

[7] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object level grouping for video shots," *Int. J. Computer Vision*, vol. 67, no. 2, pp. 189–210, 2006.

[8] R. Unnikrishnan and M. Hebert, "Extracting scale and illuminant invariant regions through color," in *Proc. British Machine Vision Conference*, 2006.

[9] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, 1988, pp. 147–151.

[10] M. Loog and F. Lauze, "The improbability of harris interest points," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 6, pp. 1141–1147, 2010.

[11] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[12] S. M. Smith and J. M. Brady, "SUSAN - a new approach to low level image processing," *Int. J. Computer Vision*, vol. 23, no. 1, pp. 45–78, 1997.

[13] P. R. Beaudet, "Rotationally invariant image operators," in *Proc. Int. Conf. Pattern Recognition*, 1978, pp. 579–583.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[15] H. Bay, T. Tuytelaars, and L. van Gool, "SURF: Speeded up robust features," in *Proc. European Conference on Computer Vision*, vol. 3951, pp. 404–417. 2006.

[16] T. Tuytelaars and L. Van Gool, "Matching widely separated views based on affine invariant regions," *Int. J. Computer Vision*, vol. 59, no. 1, pp. 61–85, 2004.

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[18] J. Maver, "Self-similarity and points of interest," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 7, pp. 1211–1226, 2010.

[19] Z. W. Miao and X. D. Jiang, "A novel rank order LoG filter for interest point detection," in *Proc. Conf. Acoustics, Speech and Signal Processing*, 2012, pp. 937–940.

[20] R. Kimmel, C. P. Zhang, A. M. Bronstein, and M. M. Bronstein, "Are mser features really interesting?," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 33, no. 11, pp. 2316–2320, 2011.

[21] P. E. Forssen, "Maximally stable colour regions for recognition and matching," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[22] H. L. Deng, W. Zhang, E. Mortensen, T. Dietterich, and L. Shapiro, "Principal curvature-based region detector for object recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[23] J. Kim and K. Grauman, "Boundary preserving dense local regions," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 1553–1560.

[24] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 1, pp. 105–119, 2010.

[25] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[26] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," in *Proc. European Conference on Computer Vision*, pp. 228–241. 2004.

[27] Z. W. Miao and X. D. Jiang, "Interest point detection using rank order LoG filter," *Pattern Recognition*, accepted with minor revision.

[28] X. D. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, no. 3, pp. 383–394, 2008.

[29] X. D. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 5, pp. 931 –937, 2009.

[30] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 2, pp. 210–227, 2009.

[31] Z. W. Miao, W. Ji, Y. Xu, and J. Yang, "A novel ultrasonic sensing based human face recognition," in *IEEE Ultrasonics Symposium*, 2008, pp. 1873–1876.

[32] C. Geng and X. D. Jiang, "Face recognition using sift features," in *Proc. Int. Conf. Image Processing*, 2009, pp. 3313–3316.

[33] C. Geng and X. D. Jiang, "Face recognition based on the multi-scale local image structures," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2565–2575, 2011.

[34] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 6, pp. 748–763, 2002.

[35] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Second IEEE Workshop Applications of Computer Vision*, 1994, pp. 138–142.

[36] "*Georgia Tech Face Database*," http://www.anefian.com/face_reco.htm, 2007.

[37] P. J. Phillips, Hyeonjoon M., S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 10, pp. 1090 – 1104, 2000.