J Inf Process Syst, Vol.12, No.3, pp.502~510, September 2016
http://dx.doi.org/10.3745/JIPS.02.0042

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Multiple Features Video Copy Detection Algorithm Based on a SURF Descriptor

Yanyan Hou*, Xiuzhen Wang*, and Sanrong Liu*

Abstract

Considering video copy transform diversity, a multi-feature video copy detection algorithm based on a Speeded-Up Robust Features (SURF) local descriptor is proposed in this paper. Video copy coarse detection is done by an ordinal measure (OM) algorithm after the video is preprocessed. If the matching result is greater than the specified threshold, the video copy fine detection is done based on a SURF descriptor and a box filter is used to extract integral video. In order to improve video copy detection speed, the Hessian matrix trace of the SURF descriptor is used to pre-match, and dimension reduction is done to the traditional SURF feature vector for video matching. Our experimental results indicate that video copy detection precision and recall are greatly improved compared with traditional algorithms, and that our proposed multiple features algorithm has good robustness and discrimination accuracy, as it demonstrated that video detection speed was also improved.

Keywords

Local Invariant Feature, Speeded-Up Robust Features, Video Copy Detection

1. Introduction

With the development of Internet and multimedia technology, multimedia sharing has become an integral part of entertainment and social networking. However, since videos are easy to copy, change, and reprint, video copy detection algorithms are being widely studied by researchers at home and abroad. Video as an ordered image set contains a large amount of feature information, and whether a video is a copy can be judged by calculating the difference between the detection video and reference video [1]. A content-based video copy detection algorithm is with good readability and robustness, feature information is extracted from the video itself during video copy detection process, no additional information was embedded into video before video released.

Most of the research on video copy detection algorithm has been derived from the image detection algorithm, where spatial features, which include global features and local features, are extracted to form feature vectors.

Global features mainly refer to video frame color, brightness, color histogram, brightness sorting, and so on [2]. HSV color histograms are extracted as video frame features in [3], but it is insufficient for

Manuscript received May 19, 2015; first revision September 22, 2015; accepted January 19, 2016.

Corresponding Author: Yanyan Hou (34844067@qq.com)
* College of Information Science and Engineering, Zaozhuang University, Zaozhuang, China (34844067@qq.com, {xzcq2003, srl9973}@163.com)

video feature description. Frame block brightness sorting based on ordinal measure (OM) is proposed in [4]. OM has better robustness against overall attacks, such as brightness change, noise, and so on, but it has worse robustness against local attacks, such as video clips, occlusion, and so on. Local features are more robust against video labeling, occlusion, video clips, and so on. The SIFT local descriptor has good stability and noise immunity, but feature calculation is too complicated [5]. The Speeded-Up Robust Features (SURF) descriptor has better robustness against geometric attacks and computational complexity is reduced as compared to the SIFT descriptor. Considering information detection accuracy and computational complexity, we are proposing a novel video copy detection algorithm based on a combination of an OM and SURF descriptor.

2. Video Copy Detection

2.1 Video Preprocessing

Video formats and attacks are different in practical application. Our proposed video copy detection algorithm is carried out in three stages — video preprocessing, feature extraction, and feature matching. Video format, frame size, and frame rate are unified during the preprocessing stage [6]. First, video luminance frames are transformed into gray frames, the frame size is unified as 352×288, and the frame rate is unified as 15 frames per second. Second, every video is segmented into fragments. Segmentation is carried out based on the block matching algorithm for video frame blocks, which is used during the OM feature matching stage. Every video frame is divided into 4×4 blocks, and the gray mean of each block is calculated. The gray mean distance between the adjacent frame blocks is calculated and compared with the preset threshold, and if the gray mean distance is more than the threshold, fragment switching occurs [7].

In order to better characterize fragment features, key frames are obtained by the superposition of video frames and Gaussian fuzzy processing is done to video frames in order to effectively detect SURF feature points.

2.2 SURF Feature Extraction

Video copy detection accuracy has been improved in the SIFT feature extraction algorithm, but the video copy detection rate is not high. The SURF feature extraction algorithm is optimized and the Gaussian two-order differential template of the Determinant of Hessian (DOH) is simplified. The filtering process is completed by addition and subtraction combination [8]. Integral frames are used to calculate feature vectors, $x(i,j)$ is the gray value of the (i,j) pixels, $y(i,j)$ is the diagonal gray integral value from the upper left corner pixel to the (i,j) pixels.

$$y(i,j) = \sum_{m\leq i,n\leq j} x(m,n) \tag{1}$$

$y(i,j)$ is obtained through two iterations of $y(i,-1)=0$ an $dy(-1,j)=0$, and $r(i, j)$ is the gray column's integral value. An all gray integral value $y(i,j)$ of a frame is composed of a corresponding integral frame.

$$r(i,j) = r(i,j-1) + x(i,j) \tag{2}$$

$$y(i,j) = y(i-1,j) + r(i,j) \tag{3}$$

DOH is optimized for reducing computational complexity, and the convolution of the Gaussian two-order differential templates with video frames were computed in the previous DOH computation. The convolution operation is then converted into a box filtering operation in the improved SURF algorithm. Each box filter operation template consists of several rectangular areas, and the filter operation template is selected according to different convolution needs. Fig. 1 shows three box filter templates. The different color areas represent different filtering coefficients: white represents the filter coefficient of 1, black represents the filter coefficient of -1, and grey represents the filter coefficient of 0.

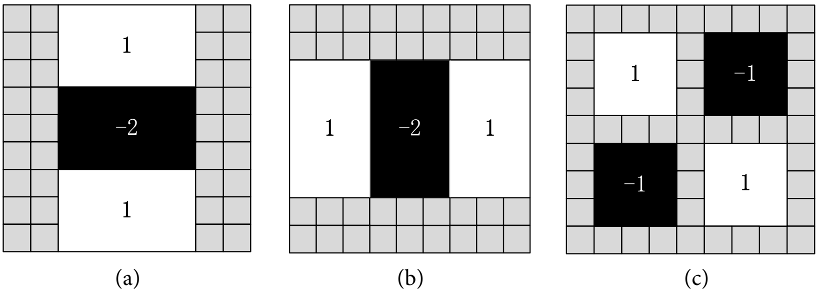


Fig. 1. Box filtering template figure. (a) x direction, (b) y direction, and (c) diagonal direction.

Video frames are filtered using the Gaussian two-order differential template, σ=1.2, and the minimum box filter template scale is 9×9. The DOH is shown in Eq. (4), and the DOH is transformed into box filtering, as shown in Eq. (5).

$$Det(H) = L_{xx}L_{yy} - L_{xy}L_{xy} \tag{4}$$

$$Det(H^{'})= D_{xx}D_{yy} - (0.9D_{xy})^2 \tag{5}$$

$v^n \in \{1,-1,-2\}$ is a vector consisting of a box filter coefficient; $\{p_1^n,\ p_2^n,\ p_3^n,\ p_4^n\}$ is a vector consisting of four integral frame vertices; S_{xx}, S_{yy}, and S_{xy} are areas corresponding to different box filter coefficients; and D_{xx}, D_{yy}, and D_{xy} are response values which corresponding to feature point.

$$D_{xx} = \frac{1}{S_{xx}} \sum_{n=1}^{3} v^n \left(p_4^n - p_2^n - p_3^n + p_1^n\right) \tag{6}$$

$$D_{yy} = \frac{1}{S_{yy}} \sum_{n=1}^{3} v^n \left(p_4^n - p_2^n - p_3^n + p_1^n\right) \tag{7}$$

$$D_{xy} = \frac{1}{S_{xy}} \sum_{n=1}^{3} v^n \left(p_4^n - p_2^n - p_3^n + p_1^n\right) \tag{8}$$

Different box filtering templates are used to calculate multi-scale feature responses. The box filtering template scales are 9×9, 15×15, 21×21, 27×27, and so on, and the integral video frames are filtered by continuously increasing the box filter template scale. Video frame response values are obtained by calculating the DOH, 3D non-maxima suppression is applied to video frame responses and different scale feature points are searched. And then, finally, the maximum response value of the feature point is searched by different filter scales [9].

2.3 SURF Feature Descriptor

In order to ensure the rotation invariance of the feature descriptor, integration frames are used to calculate SURF feature descriptors [10]. Taking the SURF feature point as the center and 6 pixels as the radius, the feature point response is calculated by the Harr wavelet, and then Gaussian weighting is applied to the feature point response value. Finally, the sliding window ω is a sector region, which uses the SURF feature point as the center, 1/6π as the angle, and 0.2 radian as the step. The Harr wavelet response dx and dy in the sliding window are accumulated to form feature vectors, and the direction corresponding to the maximum accumulation is the main direction for the feature descriptor.

$$m_w = \sum_w dx + \sum_w dy \tag{9}$$

$$\theta_w = arctan\ (\sum_w dx / \sum_w dy) \tag{10}$$

$$\theta = \theta_w | max\{m_w\}| \tag{11}$$

In order to reduce computational complexity, the frame block size is set as 5×5, the frame block response is calculated by the Harr wavelet, and the main direction weight dy and the vertical main direction weight dx are then calculated. Taking the feature point as the center, the Gaussian weighted (σ=3.3) is used to calculate dx and dy. The block response values, $\sum dx$, $\sum dy$, $\sum d|x|$, and $\sum d|y|$ are calculated to form feature vector v, which is four-dimensional and has a rotation invariance.

$$v = [\sum dx, \sum d|x|, \sum dy, \sum d|y|] \tag{12}$$

3. Video Feature Matching

In order to quickly match video features, video feature matching is completed in two phases. In the first phase, the OM features are used to match coarse video frames, the gray block mean is calculated and sorted in each key frame, the gray mean sorts the differences between the detection video frames, and then the reference video frames are calculated. The detection video is as a copy of the reference video if the difference is less than the specified threshold; otherwise the fine phase commences. In the second phase, SURF descriptors are used to match coarse video frames, and a Hessian matrix trace is considered to be the sign of a Laplasse response. During the feature point matching process, the signs of the Laplasse response are matched. If the signs of the Laplasse response are inconsistent, feature point matching is stopped, and the feature point is not viewed as a copy of the detection video feature point. However, if the signs of the Laplasse response are consistent, the feature vector v and feature angle θ will be matched.

In order to quickly search for the feature matching point, it is searched for in the corresponding frame block. If the Euclidean distance $l_{v1,v2}$ between feature vectors is less than the specified threshold λ (λ=0.5) and the feature angle difference $d\theta$ is less than the specified threshold φ(φ=10°), the feature point is as a copy of the detection video feature point.

$$l_{v1,v2} = \sqrt{(\sum dx_1 - \sum dx_2)^2 + (\sum d|x_1| - \sum d|x_2|)^2 + (\sum dy_1 - \sum dy_2)^2 + (\sum d|y_1| - \sum d|y_{2|})^2} \tag{13}$$

$$d\theta = arctan\,(\sum_w dx \,/\, \sum_w dy) \tag{14}$$

Video feature points are matched using the SIFT algorithm [11], the CGO algorithm [12], the Harris algorithm [13], and the proposed SURF algorithm. Fig. 2 is copy matching effect figure for a BUS video (the 18th frame), which is shifted. Fig. 3 is copy matching effect figure for a Football video (the 107th frame), which is rotated for different algorithms. As shown in Figs. 2 and 3, most feature points can be matched accurately under geometrical attacks in our proposed SURF algorithm.

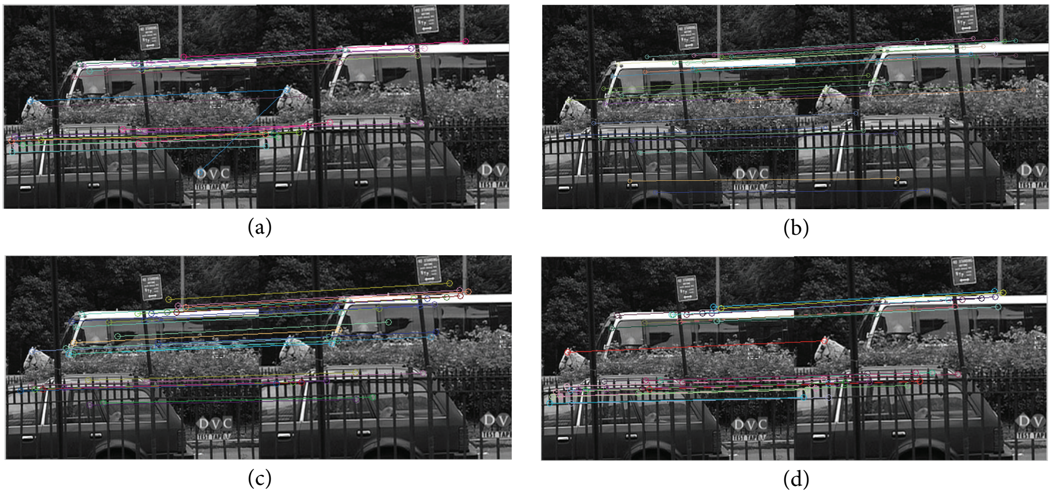


Fig. 2. Feature point matching effect figure for BUS video (the 18th frame). (a) SIFT algorithm, (b) CGO algorithm, (c) Harris algorithm, and (d) SURF algorithm.

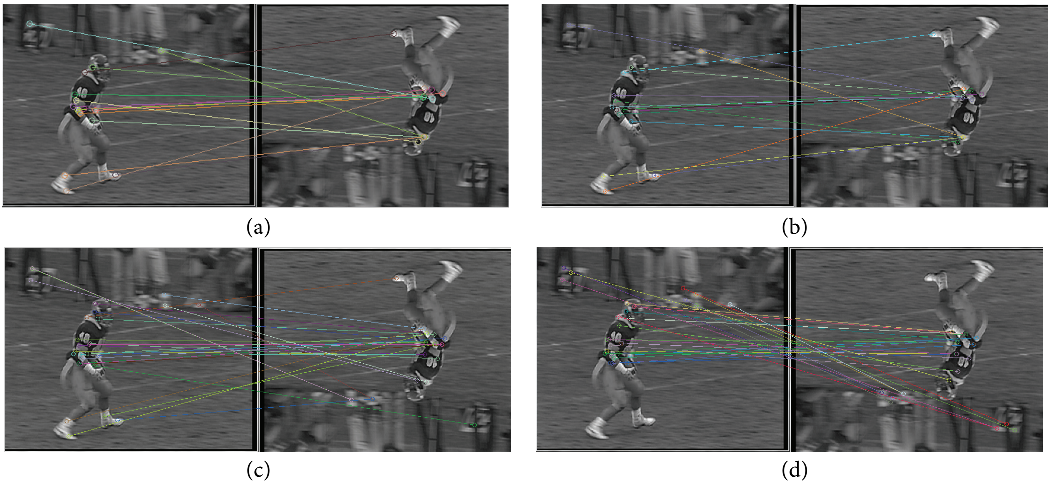


Fig. 3. Feature point matching effect figure for Football video (the 107th frame). (a) SIFT algorithm, (b) CGO algorithm, (c) Harris algorithm, and (d) SURF algorithm.

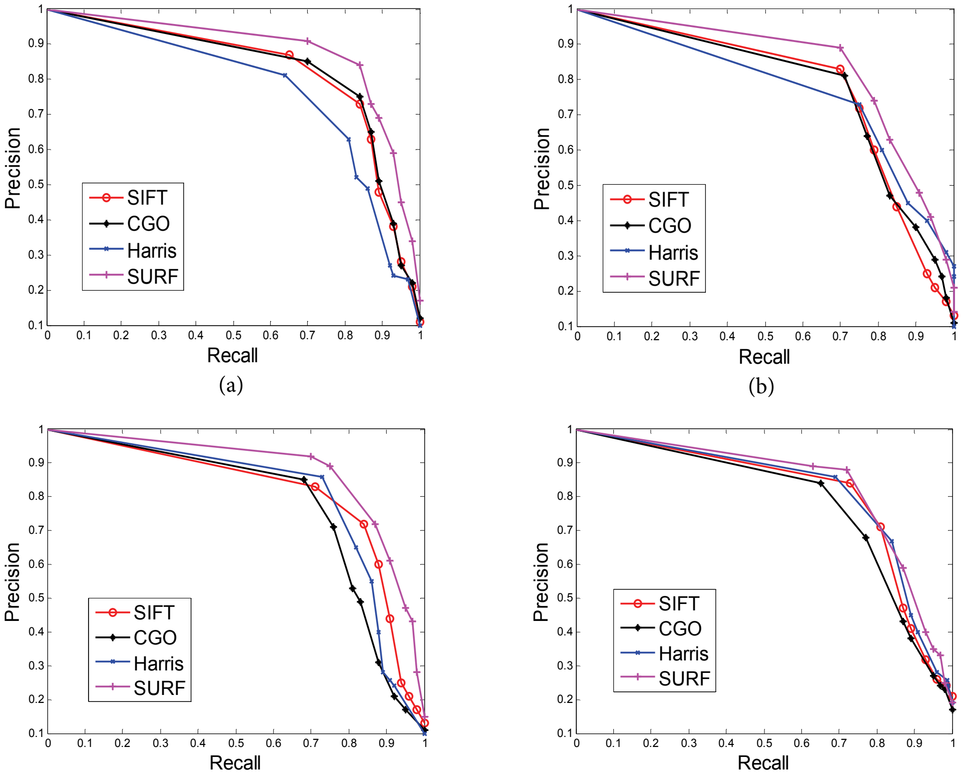
4. Experiments

In our paper, video detection accuracy, robustness, and detection time are used to verify algorithm performance. BOR11_007, Wth_03, Moon009, Anni002, etc., were selected as detection videos, and 100 random videos were selected as reference videos from the TRECVID2010 video database. TP is the number of video contents that were relevant to the source videos, FP is the number of video contents that were irrelevant to the source videos, and FN is the number of remaining videos in addition to the relevant videos. Video recall and precision were used as the evaluation standards for video detection accuracy.

$$Recall = \frac{TP}{TP+FN} \tag{15}$$

$$Precision = \frac{TP}{TP+FP} \tag{16}$$

The Precision-Recall curves reflect video detection accuracy and robustness. Video copy detection has great robustness if the precision is slowly reduced when the recall is improved. In order to further verify the accuracy of our algorithm, white Gaussian noise, frame cuts, frame shifts, and so on were used to attack videos. The comparisons of the Precision-Recall curves are shown in Fig. 4.



(a)

(b)

(c)

(d)

Fig. 4. Precision-Recall curve under several attacks. (a) White Gaussian noise, (b) frame cut, (c) frame shift, and (d) scale change.

Video detection distinction and robustness can be evaluated more accurately by *F* operator.

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (16)$$

The *F* operator is used to evaluate detection, distinction, and robustness under geometrical attacks for several algorithms. The experimental results for this are shown in Table 1. Compared with previous algorithms, video detection distinction and robustness were improved in regards to scale change, frame cuts, frame shifts, white Gaussian noise and other attacks.

Table 1. *F* operator comparison

Attack type	SIFT	CGO	Harris	SURF
Scale change	0.848	0.909	0.676	0.931
Frame lost	0.843	0.905	0.537	0.932
Frame rotation	0.821	0.843	0.715	0.868
White Gaussian noise	0.571	0.531	0.691	0.713
Frame cuts	0.883	0.914	0.829	0.936
Frame shift	0.845	0.781	0.736	0.877
Blurring	0.734	0.692	0.654	0.803
Frame drop	0.646	0.683	0.691	0.786

In order to accurately calculate video copy detection time, the video database indexing establishment time, video decoding time, and attacking time were not included in the video copy detection time. Experiments were carried out on a CPU with an Intel Core i5-4460, 8 GB memory, Windows 7 operating system, and MATLAB R2010b and VC2010 software tools were used. Three hundred different detection videos were selected as the video database, including news, sports, landscapes, and so on. The time of each video ranged from 10 seconds to 10 minutes. BOR11_007 and other seven videos were selected as detection videos. Video resolution was 352×288 and videos were in an AVI format. Video detection time comparisons with previous algorithms are shown in Table 2, from which it can be seen that the detection time was shorter and detection speed was quicker than with the algorithms used in [11-13].

Table 2. Video detection time (s) comparison

Attack type	SIFT	CGO	Harris	SURF
BOR11_007	196.9	192.4	170.1	167.2
Wth_03	185.2	189.7	173.8	141.8
Moon009	156.9	152.4	145.2	127.2
Anni002	141.7	146.8	137.3	124.1
Indi001	123.8	135.3	131.2	115.3
Boro2_010	142.3	142.1	148.1	138.2
UGS02_002	173.2	163.4	127.4	112.8
Win001001	154.6	149.8	143.7	134.6

5. Conclusions

This paper proposes a novel and effective video copy detection algorithm based on OM features and a SURF descriptor. Multi-features are used to improve video detection performance, video copy transform diversity, and real-time detection requirement are satisfied. First, the OM algorithm is used to video coarse detection. Second, the SURF descriptor is used to video fine detection. Compared with previous video copy detection algorithms based on local descriptors, video detection recall and precision are improved. Also, our proposed algorithm showed itself to be robust against most attacks, which results in the video detection real-time requirement being met. In the future, we will run this scheme under other types of modifications and improve its robustness.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China (No. 41204025). The authors would like to thank Professor Wang Hong Jun of Shandong University for the many valuable discussions concerning this work. The authors are also grateful to Professor Li Qing Hua of QILU University of Technology for their assistance with our experiments.

References

[1] S. R. Shinde and G. G. Chiddarwar, "Recent advances in content based video copy detection," in *Proceedings of International Conference on Pervasive Computing (ICPC)*, Istanbul, Turkey, 2015, pp. 346-351.

[2] S. Ozkan, E. Esen, and G. B. Akar, "Visual group binary signature for video copy detection," in *Proceedings of 22nd International Conference on Pattern Recognition (ICPR)*, Stockholm, Sweden, 2014, pp. 3945-3950.

[3] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1092-1104, 2012.

[4] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615-1630, 2005.

[5] Y. W. Zhao, B. C. Li, T. Q. Peng, and H. L. Gao, "An object retrieval method based on randomized visual dictionaries and query expansion," *Journal of Electronics & Information Technology*, vol. 34, no. 5, pp. 1154-1161, 2012.

[6] A. Joly, O. Buisson, and C. Frelicot, "Content based copy retrieval using distortion-based probabilistic similarity search," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 293-306, 2007.

[7] Y. Zhou, S. Yan, and T. S. Huang, "Detecting anomaly in videos from trajectory similarity analysis," in *Proceedings of IEEE International Conference on Multimedia and Expo*, Beijing, China, 2007, pp. 1087-1090.

[8] J. Cruz-Mota, I. Bogdanova, B. Paquier, M. Bierlaire, and J. P. Thiran, "Scale invariant feature transform on the sphere: theory and applications," *International Journal of Computer Vision*, vol. 98, no. 2, pp. 217-241, 2012.

[9] G. Mantena and X. Anguera, "Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013, pp. 8515-8519.

[10] H. Peng, C. Deng, L. An, X. Gao, and D. Tao, "Learning to multimodal hash for robust video copy detection," in *Proceedings of International Conference on Image Processing*, Melbourne, Australia, 2013, pp. 4482-4486.

[11] H. Liu, H. Lu, and X. Xue, "SVD-SIFT for web near-duplicate image detection," in *Proceedings of IEEE International Conference on Image processing*, Hong Kong, 2010, pp. 1445-1448.

[12] S. Lee and C. D. Yoo, "Robust video fingerprinting for content-based video identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 7, pp. 938-988, 2008.

[13] A. Natsev, M. Hill, and J. R. Smith, "Design and evaluation of an effective and efficient video copy detection system," in *Proceedings of IEEE International Conference on Multimedia and Expo*, Singapore, 2010, pp. 1353-1358.

Yanyan Hou

She received B.S. and M.S. degrees in School of Information Science and Engineering from Shandong University in 2004 and 2007, respectively. Since July 2007, she is as a lecturer in College of Information Science and Engineering, Zaozhuang University. Her current research interests include video communication and processing.

Xiuzhen Wang

She received Ph.D. degree in School of Information Science Technology from East China Normal University in 2009. Her current research interests include wireless communication technology.

Sanrong Liu

She received M.S. degree in School of Computer Science and Software Engineering from East China Normal University in 2007. Her current research interests include video communication technology.