Multimed Tools Appl (2016) 75:11603–11613
DOI 10.1007/s11042-015-2698-y

Human fall detection in surveillance video based on PCANet

Shengke Wang[1] · **Long Chen**[1] · **Zixi Zhou**[1] · **Xin Sun**[1] · **Junyu Dong**[1]

Received: 2 February 2015 / Revised: 13 May 2015 / Accepted: 18 May 2015 /
Published online: 12 June 2015

Abstract Fall incidents have been reported as the second most common cause of death, especially for elderly people. Human fall detection is necessary in smart home healthcare systems. Recently various fall detection approaches have been proposed., among which computer vision based approaches offer a promising and effective way. In this paper, we proposed a new framework for fall detection based on automatic feature learning methods. First, the extracted frames, including human from video sequences of different views, form the training set. Then, a PCANet model is trained by using all samples to predict the label of every frame. Because a fall behavior is contained in many continuous frames, the reliable fall detection should not only analyze one frame but also a video sequence. Based on the prediction result of the trained PCANet model for each frame, an action model is further obtained by SVM with the predicted labels of frames in video sequences. Experiments show that the proposed method achieved reliable results compared with other commonly used methods based on the multiple cameras fall dataset, and a better result is further achieved in our dataset which contains more training samples.

Keywords Fall detection · PCANet · Behavior analysis · Patient monitoring · Visual surveillance

1 Introduction

With the fast growing population of seniors, more and more elderly people in developed countries are living alone [23]. Each year, one out of three adults aged 65 years and older fall due to various reasons [10]. Fall at home is one of the major risks for elderly people, so an immediate alarming and helping is essential to reduce the rate of morbidity and mortality [20]. Since fall detection is one of the most important healthcare issues for elderly person at home, a number of research projects have been conducted for automatically detecting falls. An excellent survey on the recent developments can be found in [22].

✉ Junyu Dong
dongjunyu@ouc.edu.cn

[1] Ocean University of China, Qingdao, China

Fall detection methods are normally divided into two categories: sensor based and vision-based methods. Sensor based approaches, including wearable devices and ambient device, are widely used in the past years to detect human fall. Wearable sensor-based devices, such as accelerators and gyroscope sensors, are attached to the human body. Those wearable devices collect and provide data to a computer system or an embedded system which then analyzes the data to detect fall [5, 12, 15]. Wearable sensor-based methods have advantages of lower computational efficiency and easier to be installed. However, it intrusive and often burdensome which make elderly person do not like and forget to wear those vital sensors. Ambient devices are sensors installed on the elders' active regions [1], such as vibration sensors on the floor, to improve the performance [6, 13]. Ambient sensor based approaches eliminate the need for wear sensors all the time, however, sensors are required to be installed all over the environment. Moreover the major problem is that ambient sensor scan often generates false alarms and leads to a low detection accuracy [14]. Recent years, vision-based methods are becoming more and more popular which can be used to effectively detect multiple events simultaneously with less intrusion.

In this paper, a new method is proposed for fall detection based on computer vision and machine learning. A simple and effective deep learning network PCANet is introduced to train a single-frame human fall detection model which will be used to label every image. The structure of this paper is as follows. In Section 2, we consider the related work, whereas in Section 3, we present our human fall detection framework, which includes two stages. More in detail, Section 3.1 describes the features learning algorithm with PCANet and Section 3.2 describes the fall action detection model with SVM. In Section 4, we evaluate the performance of our model on a publically available dataset [2] and also on the dataset collected by ourselves. In Section 5, we give the final conclusions.

2 Related works

With the advances in computer vision in the last few decades, computer vision-based methods provide a promising way for detecting falls. Analysis of the person's bounding box is a simple method to detecting a fall from the surveillance video and easy to implement [18, 19]. However, a bounding box could not efficiently discriminate fall down events from fall-like events, such as sit, squat or other normal behaviors. The accuracy of this method highly depends on the relative position of the person and the field of view of the camera, and it works effectively only when the surveillance camera is placed sideways or at the same level as the human object. To better represent human shape, ellipse shape, instead of a bounding box, was introduced and good accuracy was achieved in fall detection. Chen et al.[8] used ellipse shape represented the human in the video and combined human shape analysis with other analyses, namely motion analysis and posture estimation analysis in their approaches to detect falls. In comparison with the bounding box method, ellipse shape-based approach gives a better representation of human shape and good accuracy in fall detection. However, fall-like events was also falsely detected as fall-down events using ellipses.

3 Our approach

In our approach, we apply feature learning methods to detect a fall action. The flow chart of our fall detection framework based on PCANet is shown in Fig. 1. We obtain two models after training stage: one is a single frame detection model and another is an action model. In training stage 1, the training set is extracted from video sequences with different views. Here the training samples are

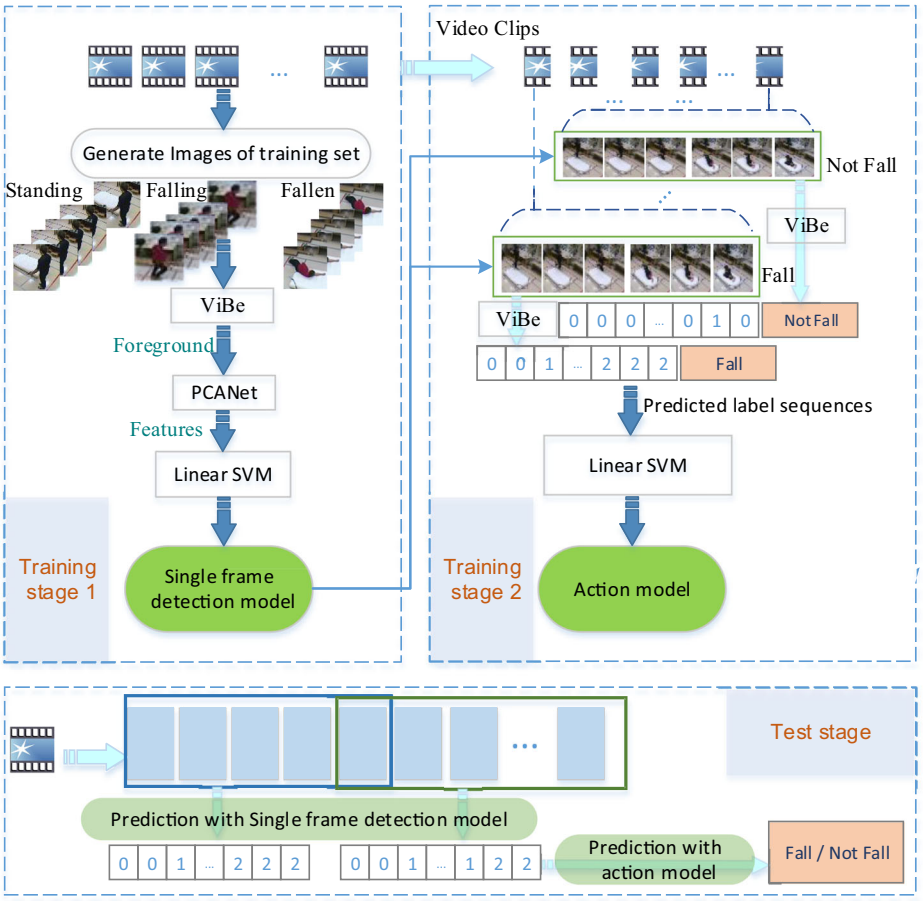


Fig. 1 Flow chart of human fall detection framework based on PCANet

formed by the sub-image including human in a specific resolution extracted using ViBe [3], which is a powerful technique for background detection and subtraction in video sequences. The training samples for the single frame detection model are labeled to three classes, namely Standing, Falling and Fallen, and then a PCANet model is trained by all the samples. Fall incident is a consistent event including many frames, so reliable fall detection should analyze a video sequence, instead of single frame. In training stage 2, based on the prediction result of the PCANet model for each frame, an action model is obtained by SVM with the predicted labels of frames in sub video clips.

3.1 Features extraction with PCANet

For image classification tasks, the hand-crafted low-level features can generally work well when dealing with some specific tasks or data process, like SIFT and HOG for object recognition [4, 11]. Yet, they are not universal for all conditions. Therefore, the thought of learning features from data of interest is proposed to make up for the limitation of hand-crafted features, and deep learning is treated as a better method to abstract high-level features which provide more invariance to intra-class variability. Compared with convolutional neural networks (CNN), a complex

architecture of neural network, which requires some expertise of parameter tuning and long-time training, PCANet is superior for its easy training and better adaption to different conditions.

The basic architecture of PCANet can be seen above in Fig. 2. It is composed of three stages: the first two stages of PCA and the last stage of hashing and histogram. Assume that we have N training images of size $m \times n$. In each image, we take a patch of size $k_1 \times k_2$ around each pixel illustrated in Fig. 3, collect all the patches, vectorize them and combine them into a matrix of $k_1 \times k_2$ rows and $(m-k_1+1) \times (n-k_2+1)$ columns.

For the ith image I_i, we obtain a matrix X_i, and we subtract patch mean from each patch and obtain:

$$\mathrm{X} = \left[\overline{X_1}, \overline{X_2}, \ldots,\right] \in R^{k_1 k_2 \times Nc} \tag{1}$$

where c denoting the number of rows of X_i. Then, we move on to obtain the eigenvectors of XX^T, and save the ones corresponding to the L_1 largest eigenvalues as the PCA filters, which can be expressed as:

$$W_l^1 = q_l\left(XX^T\right) \in R^{k_1 k_2}, l = 1, 2, \ldots, L_1 \tag{2}$$

The leading principal eigenvectors capture the main variation of all the mean-removed training patches. Thus, we finish the first stage.

At the second stage, we share similar process with stage 1. The input images I_i^l of stage 2 should be:

$$I_i^l = I_i * W_l^1, i = 1, 2, \ldots, N \tag{3}$$

the boundary of I_i is zero-padded so that I_i^l have the same size of I_i. We collect all the patches of I_i^l, subtract patch mean from each patch and obtain:

$$Y^l = \left[\overline{Y_1^l}, \overline{Y_2^l}, \ldots, \overline{Y_N^l}\right] \in R^{k_1 k_2 \times Nc}, l = 1, 2, \ldots L_1 \tag{4}$$

and, we combine the Y^l together as a matrix:

$$Y = \left[Y^1, Y^2, \ldots, Y^{L_1}\right] \in R^{k_1 k_2 \times L_1 Nc} \tag{5}$$

After that, we obtain the eigen vectors of YY^T, saving the ones corresponding to the L_2 largest eigenvalues as the PCA filters of the second stage

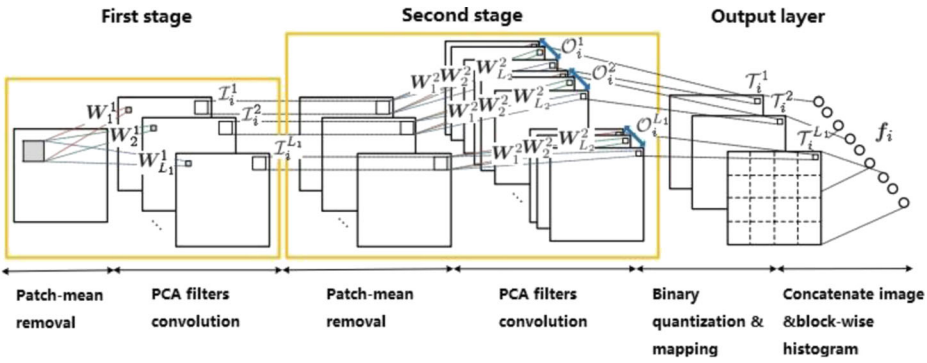


Fig. 2 The structure of the two-stage PCANet

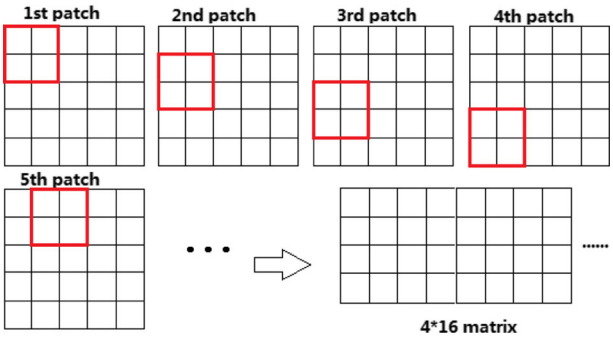


Fig. 3 Illustration of patch taking for a 5*5 image

$$W_\ell^2 = q_\ell\left(YY^T\right) \in R^{k_1 k_2}, \ell = 1, 2, \ldots, L_2 \tag{6}$$

At the final stage, for each input image of stage2, we get

$$T_i^l = \sum_{\ell=1}^{L_2} 2^{\ell-1} H\left(I_i^l * W_\ell^2\right), l = 1, 2, \ldots L_1 \tag{7}$$

The function $H(.)$ binaries these outputs, i.e. the value of the function is one for positive inputs and zero otherwise. For each of the L_1 images T_i^l, l=1,2,3,…,L_1 we partition it into B blocks, whose size is $k_1 k_2 \times B$, and we compute the $2^{L_2} \times B$ histogram matrix in each block ranging from $\left[0, 2^{L_2}-1\right]$, followed by vectorizing the matrix into a row vector $Bhist(T_i^l)$. Finally, we concatenate the $Bhist(T_i^l)$ of T_i^l, l=1,2,3,…,L_1 as the feature

$$f_i = \left[Bhist\left(T_i^1\right), \ldots, Bhist\left(T_i^{L_1}\right)\right]^T \in R^{\left(2^{L_2}\right) L_1 B} \tag{8}$$

According to specific application, the local blocks can be either overlapping or non-overlapping, and in our work, we keep it as non-overlapping [7].

As we use the PCANet to extract the features of the images of falling down, we take a set of frames of standing and fallen people from the video recording a person's falling down, and label the standing frames with 0 and fallen frames with 1. The model parameters of PCANet include the patch size k_1,k_2, the filters number L_1,L_2, the number of stages and the block size for histograms. In our experiments, we set our image size 60×60, and we set the patch size 10×10, the stage number 2, L_1=L_2=10, and the block size 25×25. The extracted features from PCANet then are put into SVM for classification with the attached labels.

3.2 Fall action detection with SVM

In our work, we trained two linear SVM models. The single-frame SVM model is used to predict the labels of individual frames generated from video sequences while the action model is to predict the label of a video.The single frames are normally classified to two states: "Fall" and "Not Fall", but for some frames it is hard to tell the label. For example, falling ends on the knees is a confused case, because in this case people is still able to move, i.e. they would stand up, or would consequently lie down. Thus, here we use three states for a single frame: Standing, Falling and Fallen. The frame with human that is clearly standing up has the state

of “Standing”, and the frame with human that is completely fallen down has the state of “Fallen”, and the one between fallen down and standing has the state of “Falling” (See Table 1). Then the features obtained from the PCANet model can be fed into a linear SVM classifier to train a single frame fall detection model.

Considering the fact that the fall behavior is a consistent action including many frames, the reliable fall detection should analyze a video sequence but not just a frame. Thus, we take every 30 consecutive frames that can capture a complete fall incident as a sub video sequence. The sub video sequences for training are classified to two states: “Fall” and “Not Fall”. We use the single-frame SVM model to predict the frames of each sub video, and we will obtain a sequence with 30 prediction labels. Combined with the sub video’s state “Fall” or “Not Fall”, each sub video can produce a training sample. The prediction results of all the sub videos form the training set to train the second stage SVM classifier.

4 Experiments

To evaluate the performance of the proposed method, we apply our model on a publicly available Multiple Cameras Fall Dataset and a dataset collected by ourselves . The previous dataset recorded simulated falls and normal daily activities from eight cameras that are mounted on the walls and in the oblique settings at 30 fps of 720×480 pixels. It consists of 24 kinds of fall incidents (11 crouching, 9 sitting and 4 lying on a sofa) from eight cameras. We extensively evaluate the proposed method on a dataset collected by ourselves. Our dataset includes 192 fall videos from four cameras mounted on the walls and in the oblique settings at 30 fps of 352×288 pixels. In the progress of collect fall videos, we choose four different directions simulating fall activities. We simulated fall activities in four different postures in every direction. For every posture we simulate 3 times fall activity.

We carry out eight experiments on the multiple cameras fall dataset and four experiments on our dataset, and one experiment tests the videos captured from the same camera view. For each experiment, we circularly choose one video as a test set at one time, and the rest of videos as training set of single frame model. The multiple cameras fall dataset record 24 kinds of short videos of fall and normal activities, and our dataset record 48 kinds of fall events. Table 1 shows the image

Table 1 Image samples of three kinds of states of single frame in Multiple Cameras Fall Dataset and our dataset

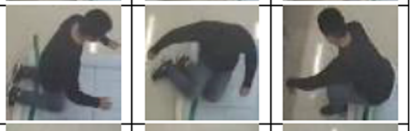
States (Label)	Multiple Cameras Fall Dataset	Our dataset
Standing (0)		
Falling (1)		
Fallen (2)		

Table 2 A sub video sequence of 30 frames to train action model in second stage. Second row is the original image sequence; third row is the foreground mask; the forth row is sub foreground image; the last row is the predicted label of single frame detection model

Frame NO.	1	2	3	…	28	29	30
Original image				…			
Foreground mask				…			
Sub foreground image				…			
Predicted label	0	0	1	…	1	2	2

samples of three kinds of states of single frame in multiple cameras fall dataset and our dataset. The three states are “Standing”, “Falling” and “Fallen” and the corresponding labels are 0, 1, and 2.

The procedure of generating the training set and the test set is shown in Table 2. First, the background model is obtained with ViBe, and then the foreground mask including human is extracted. After some morphological operations and connected component analysis, we can locate human in the foreground mask image by a bounding box. With a serial of post-process operations we can get the sub foreground images. Then the images are normalized to a resolution of 60×60 pixels to be predicted in PCANet model. The last row of Table 2 shows the predicted label.

We used sensitivity and specificity, which were commonly used in the fall detection literature [13], as indices to evaluate our proposed method.

$$\text{Sensitivity}: S_1 = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{9}$$

$$\text{Sensitivity}: S_2 = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{10}$$

where:

- TP is the number of falls correctly predicted by the system. It corresponds to the value in the first row and the first column of a confusion matrix.
- FN is the number of actual falls missed by the system

Table 3 Comparison of the proposed method with other state-of-art approaches

	Our method	Bounding box ratio approach	Chen’s approach	Chua’s approach	MHI based approach	Biomechanics approach
Sensitivity (%)	89.2	66.0	90.9	90.5	85.7	100.0
Specificity (%)	90.3	73.3	93.75	93.3	80.0	10.0

Table 4 Results of the proposed method on the dataset collected from four cameras in different views

	Camera1	Camera2	Camera3	Camera4	Average
Sensitivity (%)	93.81	93.07	89.78	78.85	88.87
Specificity (%)	98.4	98.28	99.24	88.87	98.9

- FP is the number of false detections of falls.
- TN is the number of normal activities which are correctly predicted as 'not falls'.

The proposed method was compared with five state-of-the-art approaches quantitatively on the same publicly available dataset provided in [2, 17]: bounding box ratio analysis [21], Chen's approach, MHI based approach [16], Biomechanics approach and Chua's approach [9] based on three-point representation. Experiments show that the proposed method has achieved reliable results compared with other common methods on the Multiple Cameras Fall Dataset (See Table 3), Biomechanics approach can achieve 100 % sensitivity and specificity, but people need to wear burden sensors all the time. The experiments result in our dataset achieves the 93.81 % sensitivity and 98.4 % specificity in camera 1, and it is higher than other vision-based methods (See Table 4). The average performance of 88.87 % sensitivity and 98.9 % specificity is also better than other methods in general.

5 Conclusions

In this paper, we proposed a new framework for fall detection based on automatically feature learning methods. We extracted features with PCANet, and trained two SVM classifiers to detect fall incidents. With the two models, the experiments show that our proposed method can achieve a nearly equal performance in the Multi-Camera Fall Dataset compared with other state-of-the-art methods. Moreover, we obtained a better performance when increasing the training samples in our dataset. We believe the performance can be further improved if a larger scale dataset is used and more number of camera views is involved.

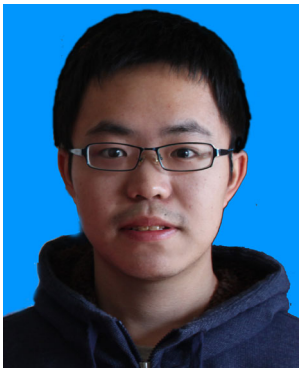
Acknowledgments This work is supported by the National Natural Science Foundation of China (NSFC) Grants 61301241, 61401413, 61403353 and 61271405.

References

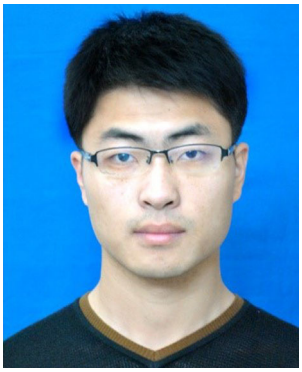
1. Alwan M et al. (2006) A smart and passive floor-vibration based fall detector for elderly. in Information and Communication Technologies, 2006. ICTTA'06. 2nd. IEEE.
2. Auvinet E et al (2010) Multiple cameras fall dataset. DIRO-Université de Montréal, Tech. Rep, 1350.
3. Barnich O, Van Droogenbroeck M (2011) ViBe: A universal background subtraction algorithm for video sequences. IEEE Trans Image Process 20(6):1709–1724
4. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Machine Intell 35(8):1798–1828
5. Bourke AK, Lyons GM (2008) A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor. Med Eng Phys 30(1):84–90
6. Cetin A (2013) Ambient assisted smart home design using vibration and PIR sensors. in Signal Processing and Communications Applications Conference (SIU), 2013 21st. IEEE

7. Chan T-H et al. (2014) PCANet: a simple deep learning baseline for image classification? arXiv preprint arXiv:1404.3606
8. Chen Y-T, Lin Y-C, Fang W-H (2010) A hybrid human fall detection scheme. in Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE
9. Chua J-L, Chang YC, Lim WK (2013) A simple vision-based fall detection technique for indoor video surveillance. Signal, Image and Video Processing: p. 1–11.
10. Hausdorff JM, Rios DA, Edelberg HK (2001) Gait variability and fall risk in community-living older adults: a 1-year prospective study. Arch Phys Med Rehabil 82(8):1050–1056
11. Hinton G, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7): 1527–1554
12. Kangas M et al (2008) Comparison of low-complexity fall detection algorithms for body attached accelerometers. Gait Posture 28(2):285–291
13. Keskin F, Töreyin BU, Çetin AE (2013) Fall detection using single-tree complex wavelet transform. Pattern Recogn Lett 34(15):1945–1952
14. Mubashir M, Shao L, Seed L (2013) A survey on fall detection: principles and approaches. Neurocomputing 100:144–152
15. Nguyen T-T, Cho M-C, Lee T-S (2009) Automatic fall detection using wearable biomedical signal measurement terminal. in Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE. IEEE
16. Rougier C et al. (2007) Fall detection from human shape and motion history using video surveillance. in Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on. IEEE.
17. Rougier C et al (2011) Robust video surveillance for fall detection based on human shape deformation. C Syst Video Technology, IEEE Transactions on 21(5):611–622
18. Tao J et al. (2005) Fall incidents detection for intelligent video surveillance. in Information, Communications and Signal Processing, 2005 Fifth International Conference on. IEEE
19. Vishwakarma V, Mandal C, Sural S (2007) Automatic detection of human fall in video, in Pattern Recognition and Machine Intelligence. Springer. p. 616–623.
20. Wild D, Nayak U, Isaacs B (1981) How dangerous are falls in old people at home? Br Med J (Clin Res Ed) 282(6260):266
21. Williams A, Ganesan D, Hanson A (2007) Aging in place: fall detection and localization in a distributed smart camera network. in Proceedings of the 15th international conference on Multimedia. ACM
22. Yu X (2008) Approaches and principles of fall detection for elderly and patient. in e-health Networking, Applications and Services, 2008. HealthCom 2008. 10th International Conference on. IEEE
23. Zambanini S, Machajdik J, Kampel M (2010) Detecting falls at homes using a network of low-resolution cameras. in Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on. IEEE.

Shengke Wang is an assistant professor in the computer science and technology department of the Ocean University of China. He acquired the B.S. degree in Computer Science from University of Jinan, China, in 2000 and received his Ph.D. in Computer Science from South China University of Technology, China, in 2005. His research interests include computer vison, machine learning, and image processing and document image analysis.



Chen Long acquired the B.S. degree in Northeast Normal University in 2013, Currently, he is pursuing a M.S. degree in Computer Architecture at the Vision Lab of Ocean University of China. His research interests are in the areas of Computer Vision and Machine Learning.



Zhou Zixi acquired the B.S. degree in electronic information engineering from Nanjing Forestry University in 2012, Currently, he is pursuing a M.S. degree in software engineering at the Vision Lab of Ocean University of China. His research interests include Computer Vision, Image Processing and Robotics.

Xin Sun is currently a lecturer at the College of Information Science and Engineering, Ocean University of China, P.R. China. He received the Ph.d degrees from the College of Computer Science and Technology, Jilin University, P.R. China, in 2013. His research interests include pattern recognition, image processing and complex network.

Junyu Dong received his B.Sc. and M.Sc. in Applied Mathematics from the Ocean University of China (formerly called Ocean University of Qingdao) in 1993 and 1999, respectively. He won the Overseas Research Scholarship and James Watt Scholarship for his PhD study in 2000 and was awarded a Ph.D. degree in Image Processing in 2003 from the School of Mathematical and Computer Sciences, Heriot-Watt University, UK. Dr. Junyu Dong joined Ocean University of China in 2004. From 2004 to 2010, Dr. Junyu Dong was an associate professor at the Department of Computer Science and Technology. He became a Professor in 2010 and is currently the Head of the Department of Computer Science and Technology. Prof. Dong was actively involved in professional activities. He has been a member of the program committee of several international conferences, including the 4th International Workshop on Texture Analysis and Synthesis (associated with ICCV2005), the 2006 British Machine Vision Conference (BMVC2006) and the 3rd International Conference on Appearance (Predicting Perceptions 2012). Prof. Dong was the Chairman of Qingdao Young Computer Science and Engineering Forum (YOCSEF Qingdao). He is a member of China Computer Federation (CCF), ACM and IEEE. Prof. Dong's research interest includes texture perception and analysis, 3D reconstruction, video analysis and underwater image processing.