Human Activity Recognition Based On 2d Texture Signal Pattern Analysis

Sathiya P, Anna University, India
AnandhaKumar P, Anna University, India

Abstract

Human activity recognition is an important research area of computer vision which dictates the need to automatically detect and retrieve semantic events in videos based on video contents. In this paper, we attempt to extract the foreground object from the video clip using color model and generate a unique signal pattern for the detected foreground (human). Signal pattern is generated for the extracted 2D texture features and the most significant features are selected using feature selection method. For each detected object, we can study its corresponding motion pattern, entry/exist points, and behavior patterns. Based on this information, it is efficient to improve the object detection and track the abnormal event occurrence. Experiments were performed on KTH dataset, High-Level Human interaction dataset and real time video dataset. The empirical results show that 85% of accuracy based on precision/recall measure was obtained, and the ability to recognize the activities in real time shows the promise for applied use.

Keywords: Human Activity Recognition, Texture Feature, 2D Texture Signal Pattern.

Introduction

In recent days, automatic human activity recognition plays an important role in the research area of computer vision. It has gained its importance in the field of video surveillance, healthcare monitoring system, interaction between human and monitoring the activities of pedestrian in traffic...etc. In all the above mentioned applications automatic recognition of complex human activities, is a combination of single or multiple human interactions which are crucial cue for activity recognition. On the other hand, recognizing human activities in a dynamic scenario is a challenging task due to its camera location, illumination changes, and variations in dynamic environment and occlusions in the frame sequences. An action is a process of performing some set of action to achieve something which leads to an activity. Hence the main objective of the system is to automatically classify the different activities performed by the humans. Humans perform different types of activities. Based on the complexity, the human activities are diversified into different levels like boxing, walking, running, jumping etc... All the above mentioned activities fall under the category of gesture classification. The main objective of the proposed system is to focus on different scenario in recognizing various human activities happening in video.

This system focus on the detection of visible low-level action primitives and actions by using wave pattern analysis. Detect the foreground objects automatically in every consecutive frames of the video in content based video retrieval system. Then wave pattern is detected for foreground objects in every frame by using texture features. This wave pattern is used to train the classifier. It should give alert about the activity of humans automatically. Different types of features are extracted in order to identify the different activities performed by humans in videos. In earlier days the different activities performed by humans were classified based on their shape, texture, color and motion. Using these extracted features it is really hard to classify the different actions performed by humans in real time scenarios due to variation in illumination, occlusion and changing background environment. In order to overcome these challenges, various feature points are extracted from the human pose to model different human actions.

Related Work

Human Activity Recognition (Gaglio, 2015) proposed 3-D posture using High-level information which are obtained from "Kinect" to learn the different activities performed by

humans. Machine learning techniques were used to identify and model the different kind of activities performed on Kinect Activity Recognition Dataset. The system was able to achieve the precision and recall of 77.3% and 76.7% respectively. The misclassification rate was high during the frame loss and body occlusion scenarios. An adaptive foreground detection method (Singh, 2008) was proposed to extort data related with various activities performed by humans. Values obtained from the directional vectors were used to group the mixture of activities carried out by humans. Time taken to recognize the different activities and to minimize the errors, vibrant characteristics is taken into consideration. The training data set has to be changed for detecting people with different body structure. The CRRs was in between 85% to 99% for eight different activities when the temporal smoothing was not applied in the detection system. But, recognition rate was about 100% when the temporal smoothing was applied in the detection system. Fuzzy models in vector quantization and linear discriminant analysis methods (Iosifidis, 2012) were applied to understand the linear combinations of patterns for representing different activities. The proposed method was able to identify the person performing different activities in various viewing angles. Centroid location plays a vital role in classifying the different activities. By using the accumulation of different centroid information the system was able to identify the person using a Bayesian approach. The recognition rate was about 97.08% using 340 dynemes. The system has to be modified if the detection has to be performed on different types of applications. Bag-of-words (Mukherjee, 2011) method was proposed which consists of large vocabulary of poses and extracts a distinct pose using centrality measure of key poses. It calculates the histogram of oriented field vectors only for a single person and this method is not suitable for key pose detection of multiple people in a particular frame. The system outperforms well for single person, activity recognition and it has to be improved for multiple person activity detection.

A context space model (Wiliem, 2012) was proposed for detecting anomalous behavior in the entire video sequence. Context information is very much useful to create a detection system. Whereas in the existing surveillance system, the use of contextual information is limited, in automatic anomalous human behavior detection. This context space model provides guidelines for the system designers to select information which can be used to describe context and also enables a system to distinguish between different contexts. Context space is defined as a n-dimensional information space formed by context parameters selected from the

contextual information as its bases. The use of contextual information in these approaches is still limited. The recognition rate is 93.4%.

Network-transmission-based algorithm (Lin, 2014) was proposed to identify the human activities in video sequences. Using the network concept, the node represents the particular scene in the video sequence and the correlations between the scenes are represented using the edge information. The movements of humans in the video sequence are related to the packages in the network and they are monitored as transmission in the network. The main drawback of the proposed system is it will stop recognizing the activity after identifying the abnormality in the video sequence and the position of the camera is fixed at a particular position. An articulated and generalized Gaussian Kernel Correlation (Ding, 2016) method was applied to detect the different kind of human pose. The Gaussian Kernel correlation was used to represent the similarity between the previous templates and they were represented using SoG variants. Shape modeling was done using the information gained from the kinematic skeleton in a multivariate SoG template.

Proposed System Overview

The main aim of the automatic surveillance system is to extract the human body region in the dynamic environment with illumination variation at an instantaneous time rate. The challenges include how to identify the human region in each frame and how to extract the specific feature which describes the human and background in real time scenario. The existing techniques cannot be incorporated into the proposed system due to its specific significance. Figure 1 shows the intact architectural view of the proposed work. Real time video is taken and it is converted into frame sequence for monitoring the continuous variation at each time instant precisely. RGB based input frames are converted into HSV color space since these models are more responsive to human eyes and they have direct impact on human perception.HSV color space based GMM model is used for subtracting the background to identify the foreground objects.

To extract the region of interest i.e. human from the background, connected component and pre-processing techniques like filtering and morphological operations were performed to get a fine tuned foreground objects. The texture pattern of the foreground objects are derived by applying transformation based method. The derived energy values are used to estimate the

various activity performed by humans. Multiclass SVM classifier is trained with large dataset of different human activities like single person actions and human-human interactions. Each and every minute action is considered to note down the difference in the activities performed by humans. Different actions performed by humans will be distinct because the energy value graph is unique for each activity.

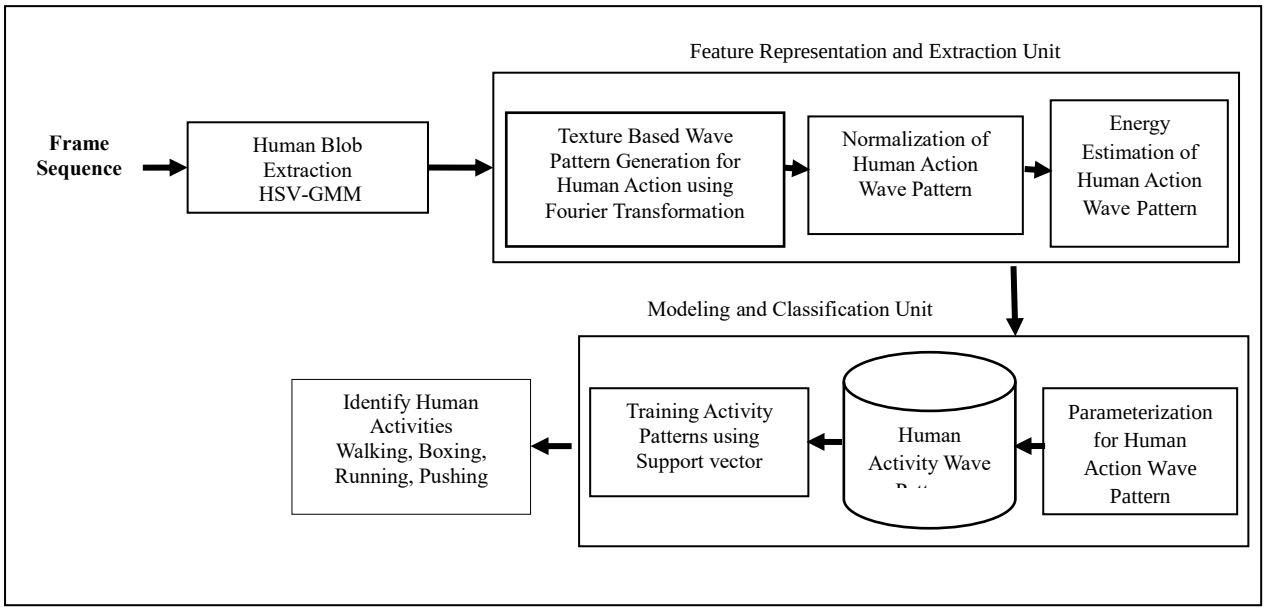


Figure 1. Proposed architecture diagram

Human Activities Training And Detection With 2d Texture Signal Pattern

In this part, an elaborated steps about the 2D texture signal pattern to differentiate the human or non-human and identification of different activities performed will be discussed. The detailed description required for training and identifying the activities are given in this section.

Transformation Based 2D Texture Pattern Description

In this paper, the texture pattern for human action is obtained using 2D Fourier transform. Let us consider a real-time color video (V) which consists of 'n' number of frames (f) denoted as f1, f2,...fn. Each video sequence is denoted as I(x,y,t) where (x,y) denotes the position co-ordinates of each pixel at particular time period t. The RGB color model, are converted into

HSV color space and Gaussian mixture model is used to extract the moving foreground object from the image sequence.The probability of the current pixel Xt is calculated by

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \sum_{i,t}) \quad (1)$$

where K is the number of distributions in the mixture, ωi,t represents the weight of Kth distribution in frame t, μi,t represents the mean of Kth distribution in frame t, ∑i,t denotes the standard deviation of Kth distribution in frame t and Η (Xt , μi,t, ∑i,t) denotes the probability density function. A threshold value will be fixed, if the current pixel value exceeds the threshold value it will be considered as a foreground object. After the extraction of foreground object, non-human objects are removed using morphological operators followed by blob analysis. Fourier transformation technique is applied on the human region H(i,j) of blob size MxN using the equation

$$F(k,l) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} H(i,j)\, e^{-i2\pi(\frac{ki}{M}+\frac{lj}{N})} \quad (2)$$

where the power spectrum obtained in frequency domain contains both phase and magnitude values. Phase values pose a smaller amount of information about the human region where as the magnitude coefficients are used to analyze the different properties of texture. The intensity value obtained by applying Fourier transformation is wide in range and logarithmic transformation is applied over it to bring down the dynamic range of human image by substituting every intensity values by its logarithm. The logarithm transformation is evaluated by using

$$L(i,j) = k \log(1 + |H(i,j)|) \quad (3)$$

where k represents the scaling constant of an intensity value with a maximum value of 255. Texture values contain the information about the irregularity, directionality and coarseness properties of human region. The energy level is more in high frequency section for the irregularity texture patterns. The magnitude value represents the direction in which the human region is moving. Motion vector values are estimated for the human region as a result of texture classification. So motion vector values are generated using magnitude vector values in each frame at a particular time t.

Feature Extraction

Motion analysis of human region is a significant thing in human activity recognition. The feature vector generated out of motion analysis identifies the region where the movement has taken place in the frame sequences at a particular time instant t. Motion descriptors are

generated by extracting the dominant feature set of different human actions bounded by bounding box in each frame sequence to describe the different kind of actions performed by humans. This is carried out by applying optical flow method on each bounding boxes in each frame sequence. Texture - Motion descriptor TM is divided into horizontal TMH and vertical THV section which carries the relevant information about the motion descriptor. The TMH and TMV motion descriptors are estimated using optical flow where $TM_H = \{x_1, x_2, \dots, x_h\}$, 'h' represents number of horizontal motion components and $TM_V = \{y_1, y_2, \dots, y_v\}$, 'v' represents number of vertical motion components. The TM feature vector $F = \{\mu_1, e_1, \dots, \mu_n, e_n\}$ are generated by computing the mean (μ) and energy (e) values of human region where n denotes the number of video frames. The features such as mean and energy are calculated for the human region of blob size I x J using the following equations

$$Mean = \frac{TM_H + TM_V}{n} \quad (4)$$

$$Energy = \sum_{s=0}^{I-1} \sum_{t=0}^{J-1} \left| H(s,t) \right|^2 \quad (5)$$

Activity Pattern Generation

The human activities can be divided into two different classes purely based on the actions they perform 1) Low - level activities and 2) High - level activities. Low - level activities deals with the actions performed by a single person like walking, running, jumping..., etc. High - level activities is an integration of several kind of single human actions. The human actions are mainly classified with selection of feature parameters taken for consideration. The energy level e_L keeps on varying for different kind of human activities. As the energy values cannot be directly taken from the frame sequences, it undergoes transformation to measure the motion vector direction with respect to Rotation, Scaling and Translation properties. Consider a human activity dataset D which comprises of single or complex type of human activities, each of which generates a set of TM feature vectors and their own energy levels e_L. The generated feature vectors are used to differentiate the infinitesimal changes between different classes of actions. For example, energy level e_L of the single person activity $Es = \{a_1, a_2, \dots, a_{single}\}$ varies with respect to the different poses of human throughout the entire video sequence. Where a_{single} represent the individual activity performed by human for instance $'a_1'$ represents walking activity and a_2 represents running activity and it goes on for KTH dataset. For UT Interaction dataset (human - human interaction) consists of several complex activities E_C and energy level e_L is given as $E_C = \{b_1, b_2, \dots, b_{complex}\}$, where

$b_{complex}$ represents the interaction between person and the energy level for such kind of interactions will be different and unique. For example, $'b_1'$ represents the handshake interaction and $'b_2,'$ represents the pushing interaction and the activities go on for UT interaction dataset.

Classifier

Classification using machine learning techniques are widely used for human activity recognition system. The classifiers are modeled using the energy level e_L pattern to identify the type of activities performed by humans. The training set $C = \{C_1, C_2\}$, where C_1 represents the human activity performed by single person and C_2 represents the human - human interaction between two or more than two persons. The multi - class SVM classifier are used to two different classes separately. Let us consider a training set $S = \{s_1, s_2, \ldots s_{single}\}$, which represents the different kind of activities performed by a single person (walking, jumping, running....) and their corresponding energy level e_L are given as $Es = \{a_1, a_2, \ldots, a_{single}\}$. When new test video are given as input to the automatic monitoring system, the class given as an output represents any one of the's' associated classes from the given equation,

$$T(S) = C_1 \text{ (argmax f (a, single))} \quad (6)$$

Similarly for UT interaction dataset (human-human interaction dataset) the training set is given as $M = \{m_1, m_2, \ldots m_{complex}\}$, where 'm' represents the different kind of interaction (complex actions) taking place in particular video sequence and their corresponding energy level is given as $E_C = \{b_1, b_2, \ldots, b_{complex}\}$. For new instances, the multi-class SVM classifies the activities based on their energy level using the following condition,

$$T(M) = C_2 \text{ (argmax f (b, complex))} \quad (7)$$

By this process the multi-class SVM was able to classify the different kind of activities performed by humans and generate a unique pattern for different type of activities.

Experimental Results

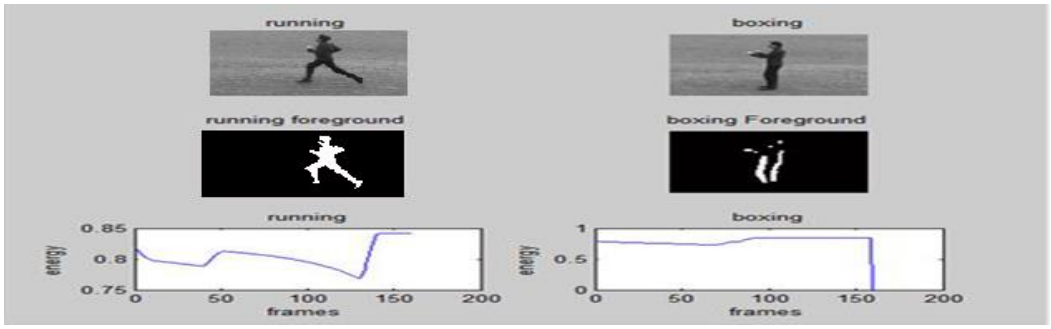
The proposed method was tested on KTH and UT Interaction datasets. The KTH dataset contains six different kind of human activities performed by a single person with different environmental conditions. In total there are 2391 sequences in the entire database at 25fps frame rate. The average time of the video sequence is about 4 sec with image resolution of

160x120 pixels. Figure 1 shows the sample from KTH dataset. The UT interaction dataset consists of six various kind of human- human interaction with the average video length of 1minute approximately. On an average there are 20 video sequences with resolution of 720x480 pixels at 30fps frame rate. The height of a person in all the video was about 200 pixels and they are taken with 15 different clothing conditions. The video consists of 10 set of video sequence taken in parking scenario with background static and small camera jitter. Figure1 shows sample video sequence from both datasets.

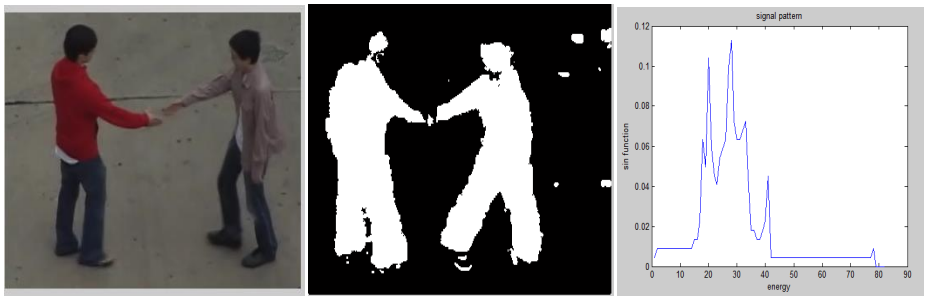
(a) KTH dataset (b) UT interaction dataset

Figure 1. Sample frame sequences

After extracting the energy level values for each human activities the wave pattern are generated to represent the movement of human and the energy spent for that particular action. In figure 2 the energy level variation are given for different human activities for both datasets and they are unique wave pattern for the particular actions. These energy values are given as input to the multi-class SVM to classify the human activity. Figure 2(a) shows the output for energy level variation of two different activities for entire video sequence of KTH dataset. Whereas, Figure 2(b) shows the energy level variations of particular activity (handshake) for the entire frame sequences.



(a)



(b)

Figure 2. (a) Energy pattern for KTH dataset (b) Energy pattern for UT interaction dataset

The multi - class SVM classifier was used to train and test the different kind of human activities performed in both KTH and UT Interaction video sequences. With the information generated from the human activity energy pattern the system was automatically able to classify the action performed in each and every frame sequences. Figure3(a) shows the result of activity classification for KTH dataset and (b) shows the result for UT Interaction classification for handshake human activity.

(a)KTH dataset (b)

UT interaction dataset

Figure 3. Multi - class SVM human activity classification

The confusion matrices for both KTH and UT Interaction dataset are shown in the Table I and Table II respectively. They are derived from the energy feature vectors which are extracted transformation based pixel information. The 2D texture signal pattern model gets an accuracy of 85. 3 % for KTH dataset and 84.374 % for UT Interaction dataset.

Table I Confusion matrices accuracy percentage of KTH dataset

	Handclap	**Boxing**	**Walking**	**Running**	**Waving**	**Jogging**
Handclap	83.33	16.67	-	-	-	-
Boxing	-	100	-	-	-	-

Walking	-	6	81	13	-	-
Running	-	-	14	76	-	10
Waving	15.4	-	-	-	84.6	-
Jogging	-	-	7.6	5.2	-	87.2

Table II Confusion matrices accuracy percentage of UT Interaction dataset

	Handshake	**Kicking**	**Pushing**	**Hugging**	**Pointing**	**Punching**
Handshake	84.76	-	15.24	-	-	-
Kicking	-	81.7	10.2	-	-	8.1
Pushing	10.15	-	82.75	7.10	-	-
Hugging	-	-	-	87.98	12.02	-
Pointing	-	-	16.46	-	83.54	-
Punching	-	14.54	-	-	-	85.46

Our proposed method shows that each and every activity performed by human will have their own energy level and they can be classified based on their unique energy patterns. The proposed system performed well and produced a better accuracy than the existing system as shown below. Table 1 and Table 2 shows the comparison of KTH and UT Interaction dataset respectively with different existing methods and their accuracy rate of classifying the human activities based on the different kind of feature points.

Table III Performance comparison of KTH dataset

Author and Year	**Feature**	**Classifier**	**Accuracy (%)**
(Schuldt, 2004)	Spatiotemporal intensity points	SVM	71.83
(Ke, 2005)	Optical Flow	Boosting	62.97
(Dollar, 2005)	Gabor Filter	1 – NN	81.17
(Oikonomopoulos, 2009)	B- Splines	Gentle Boost + RVM	80.8
PROPOSED METHOD	**Energy Points**	**Multi – Class SVM**	**85.08**

Table IV Performance comparison of UT interaction dataset

Author and Year	Feature	Classifier	Accuracy (%)
(Laptev, 2005)	Space-time interest points	Naïve Bayes	54.5
(Laptev, 2005)	Space-time interest points	SVM	65.5
(Dollar, 2005)	Sparse spatio-temporal points	Naïve Bayes	53.5
(Dollar, 2005)	Sparse spatio-temporal points	SVM	70
(Leibe, 2008)	Scale-invariant interest point	Hough Forest	77
PROPOSED METHOD	**Energy Points**	**Multi – Class SVM**	**85.08**

Conclusion

The proposed model for human activity recognition uses the energy pattern for recognizing the different kind of activities performed by KTH and UT Interaction dataset. The multi-class SVM classifier performs well and was able to improve the accuracy rate of classification. The different kind of human movements were efficiently recognized using new kind of feature points. As a future work the system will be trained to identify and recognize the different kind of human activities in dynamic environment.

References

Ding, M., & Fan, G. (2016). Articulated and Generalized Gaussian Kernel Correlation for Human Pose Estimation. IEEE Transactions on Image Processing, 25(2), 776-789.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005, October). Behavior recognition via sparse spatio-temporal features. In 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (pp. 65-72). IEEE.

Gaglio, S., Re, G. L., & Morana, M. (2015). Human activity recognition process using 3-D posture data. IEEE Transactions on Human-Machine Systems, 45(5), 586-597.

Iosifidis, A., Tefas, A., & Pitas, I. (2012). Activity-based person identification using fuzzy representation and discriminant learning. IEEE Transactions on Information Forensics and Security, 7(2), 530-542.

Ke, Y., Sukthankar, R., & Hebert, M. (2005, October). Efficient visual event detection using volumetric features. In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (Vol. 1, pp. 166-173). IEEE.

Laptev, I. (2005). On space-time interest points. International Journal of Computer Vision, 64(2-3), 107-123.

Leibe, B., Leonardis, A., & Schiele, B. (2008). Robust object detection with interleaved categorization and segmentation. International journal of computer vision, 77(1-3), 259-289.

Lin, W., Chen, Y., Wu, J., Wang, H., Sheng, B., & Li, H. (2014). A new network-based algorithm for human activity recognition in videos. IEEE Transactions on Circuits and Systems for Video Technology, 24(5), 826-841.

Mukherjee, S., Biswas, S. K., & Mukherjee, D. P. (2011). Recognizing human action at a distance in video by key poses. IEEE Transactions on Circuits and Systems for Video Technology, 21(9), 1228-1241.

Oikonomopoulos, A., Pantic, M., & Patras, I. (2009). Sparse B-spline polynomial descriptors for human activity recognition. Image and vision computing, 27(12), 1814-1825.

Schuldt, C., Laptev, I., & Caputo, B. (2004, August). Recognizing human actions: a local SVM approach. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 3, pp. 32-36). IEEE.

Singh, M., Basu, A., & Mandal, M. K. (2008). Human activity recognition based on silhouette directionality. IEEE Transactions on Circuits and Systems for Video Technology, 18(9), 1280-1292.

Wiliem, A., Madasu, V., Boles, W., & Yarlagadda, P. (2012, April). A Context Space Model for Detecting Anomalous Behaviour in Video Surveillance. In Information Technology: New Generations (ITNG), 2012 Ninth International Conference on (pp. 18-24). IEEE.