TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 15, No. 1, July 2015, pp. 162 ~ 170
DOI: 10.11591/telkomnika.v15i1.8091

Real-time Vision-based Hand Gesture Recognition Using Sift Features

Mitra khaledian*[1], Mohammad bagher Menhaj[1,2]
[1]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Qazvin, Iran
[2]Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran,
*Corresponding author, e-mail: Mitra.khaledian@yahoo.com[1], Menhaj@aut.ac.ir[2]

Abstract

This paper introduces a new algorithm based on machine vision for the recognition of hand gesture. In step 1, the Microsoft Kinect sensor is used to capture color images and depth. User's hand detected by eliminating the background and rescaling image. In the next step, "Scale-invariant feature transform (SIFT)" algorithm is used for the feature extraction. The extracted feature vectors are built in vocabulary tree with K-means clustering. Finally, Hand gesture is recognized by a recognition high-level method called stochastic context-free grammar (SCFG). SCFG is used for syntactic structure analysis that is based on hand gesture recognition, that is combined postures can be analyzed and recognized by a set of production rules. SCFG is most effective in disambiguate. By this approach, we are able to recognize various gestures in 30 frames per seconds (fps) and with more than 90 % accuracy.

Keywords*: hand gesture, human-computer interaction, vocabulary tree, stochastic context-free grammar*

1. Introduction

Gesture is a meaningful concept of complex human movements in order to make an intelligent and efficient communication between human and computer. Recognition of hand gestures is one of the important applications of image processing in recent years. The computer's ability to recognize hand gestures and analyze how it moves in different environments has many applications. The general areas of interest in this study, is human-machine interaction (HCI). The meaning of the human-machine interactions is conversations and interactions between human users on the one hand and on the other hand, are computers and intelligent software factors. Processing video images with the aim of recognizing hand gestures is one of the most attractive and most used areas of research in the field of machine vision over the past decade. The conducted researches show using of two types of hand gesture in human-computer interaction. In some systems, consider a combination of multi-mode hand postures in sequential frames equivalent to a command, while on the other systems, hand postures in each frame have independent meaning. The first kind of motion is called dynamic gestures and the second kind is called static gestures [1]. Different hand gesture recognition algorithms can be classified in two respects [2]: Appearance-based approaches (two-dimensional model) and three-dimensional model-based approaches [3]. In appearance-based approaches, a two-dimensional image feature is used to model the visual appearance of the image and these parameters utilize to compare with the input images. 3D three-dimensional model-based approaches independent to its significant freedom degree and try to obtain estimates of the hand parameters, then compare it with the input image [4].

In machine vision, there has been much research in human-machine interaction that has been conducted with different types of cameras. In [5] the single lens camera, in [6] a multi-lens camera, in [7] depth diagnosis and in [8] Infrared Cameras has been used. Different lenses give us useful information to improve recognition accuracy more and more.

One study [9] reported a method based on color of skin in the image. However, this method is very sensitive to lighting conditions and required that no other skin-like object exist in the image. In [10], using Adaboost learning algorithm and SIFT features leads to the rotation invariant hand detection. SIFT method [11] is a robust feature detection to represent image

Received February 11, 2015; Revised May 14, 2015; Accepted June 1, 2015

based on key-points. The keypoints provide rich local information of an image. However, several features such as a contrast context histogram had to be used to achieve hand gesture recognition in real time. In order to achieve real-time performance and high recognition accuracy, Juan et al. [12] evaluated performance of SIFT, principal component analysis (PCA) – SIFT, and speeded up robust features (SURF) by many experiments. SIFT algorithm extracts features, which are invariant to the rotation and scale from images. PCA-SIFT, which is introduced in [13], employs PCA to normalize gradient patch.

Here, we focus on bare hand gesture recognition without the help of any markers and gloves. To be robust against a cluttered background and various lighting conditions, we used a depth map, which contains information relating to the distance of objects from a viewpoint. For this aim, Kinect sensor is utilized to capture both the color image and its corresponding depth map. Using the depth map, hand can be accurately detected according to distinct gray-level in our test environment. The detected hand is extracted by replacing hand area with a black circle. After extracting the hand, the hand area only is saved in a small image, which will be used in extracting the features by scale invariance feature transform (SIFT) algorithm. For the first time, Lowe [11] proposed using SIFT features which are invariant to scale, orientation and partially invariant to illumination changes, and are extremely distinctive of the image. Therefore, SIFT features are extracted from the hand detected images. After this step, a vocabulary tree is offline trained by the hierarchical K-means clustering. Next, a weighted vocabulary tree using Term Frequency Inverse Document Frequency (TFIDF) weighting is built to recognize postures using k-nearest neighbor and voting. The statistical approach can quantitatively describe the hand posture using numeric parameters. However, the quantitative description is not adequate to represent a hand gesture's structural information. In this situation, a syntactic object description is more appropriate to represent the composite characters of hand gestures [14]. With a grammar-based approach to convey the hierarchical nature of hand gestures, we can construct a concrete representation for the hand gestures and, thus, enable the system to recognize the gestures based on a set of primitives and production rules.

In this paper, in Section 2, the overall structure of the system and the detection of hand and then feature extraction procedure proposed by the SIFT algorithm, is given. Section 3 describes how to build a vocabulary tree with k-mean clustering and in the next section; we describe the recognition of hand gesture and the final section summarizes and concludes the paper.

2. The Overall Structure of the System

As we said, we provide an algorithm based on machine vision that detects users' hands in real time, and for hand gesture recognition is fast and accurate. Figure 1 shows the gesture recognition process in our proposed algorithm. In this section, we analyze the process of hand gesture recognition.

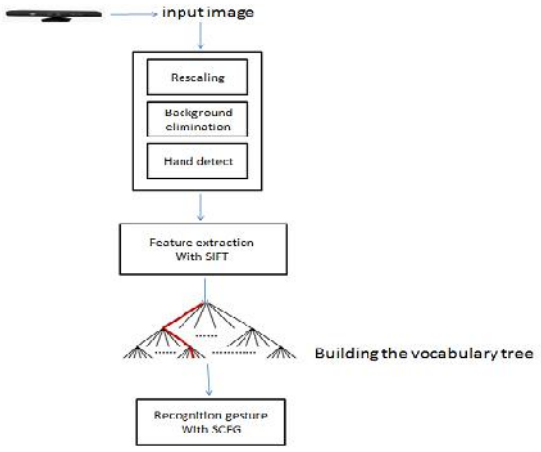

Figure 1. Illustrating overall picture of the system

2.1. Segmentation

Segmentation involves four steps, which are shown in Figure 3. As you can see, the process involves taking a photo using the Kinect camera, rescaling, background elimination and hand detection.

2.1.1. Input Image

xplaining research chronological, including research design, research procedure (in the form of algorithms, Pseudocode or other), how to test and data acquisition [1-3]. The description of the course of research should be supported references, so the explanation can be accepted scientifically [2, 4].

Tables and Figures are presented center, as shown below and cited in the manuscript.

Figure 2. Block diagram of the extraction hand

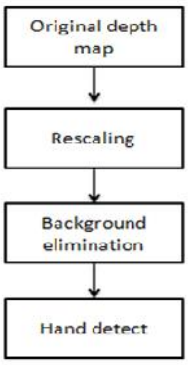

Figure 3. Block diagram of extraction Hand

2.1.2. Rescaling

In this step, rescaling is used to improve the raw image. If the user does not work properly with Kinect, the maximum depth of the points will be on his arm. With this change of scale, the range of the gray levels on hand, and the range of white levels of the body and background are displayed. Part "b" of Figure 4 shows the image obtained from this point.

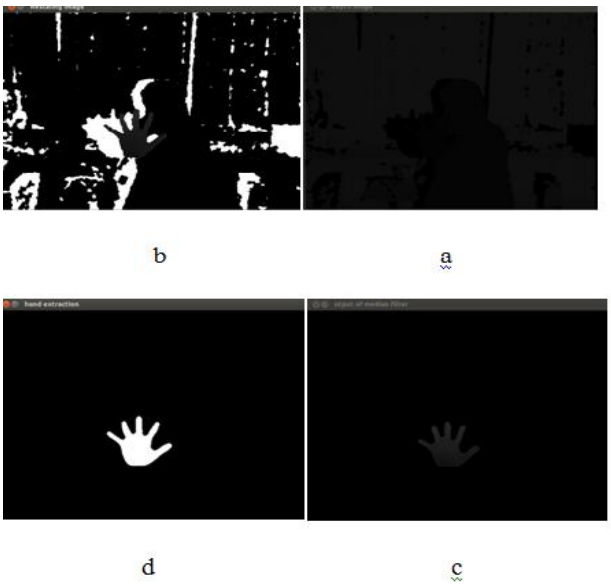

Figure 4. Segmentation of the hand

2.1.3. Background Elimination

The white background and the body are not necessary for extraction, so we performed a digital-negative operation on the image to make the image most optimal for subsequent steps. The shadows around the hand and the body, due to the positional disparity between the RGB camera and the depth sensor, can severely affect the performance of the recognition, so we have to eliminate them. Based on the previous step, the shadow at this point is white, while the background is black. The gray part is the hand, which is in the sensitive range. We can eliminate the shadows by thresholding the gray level. Part "c" of Figure 4 shows the image obtained from this point.

2.1.4. Hand Detect

We captured the top point of the palm of the hand, and kept the pixels, which represent the depth up to 24 units from that point. Otherwise, they will be eliminated. All pixels in this range will be converted into white. Finally, we extracted a hand shape image for recognition. Part "d" of Figure 4 shows the image obtained from this point.

3.Features Extraction using the Scale Invariant Feature Transform

The main features of the SIFT algorithm which motivate us to apply this algorithm, are invariant to scale and rotate and real time extraction of low-resolution images. The SIFT algorithm extract features in four stages:

First stage: A set of difference of Gaussian filters applied at different scales all over the image, and then the locations of potential interest points in the image are computed.

Second stage: The potential points are improved by removing points of low contrast. Third stage: Assigning an orientation to each key point based on local image features. Fourth stage: Computing a local feature descriptor at each key-point, which is based on the local image gradient, transformed according to the orientation of the key-point to provide orientation invariance.

The extracted feature vectors from each hand image in this step are used to train our hand gesture recognition system. The number of key-points is dependent on the area of the detected hand.

The extracted feature vectors were used to train our recognition system. Key points will depend upon the area of handwriting recognition, our training phase for the images in the database, 40% to 86 key points and for each of the key points exists of a 128-dimensional feature vector.

n this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [2, 5]. The discussion can be made in several sub-chapters.

3.1. Build Vocabulary Tree using K-means Clustering

x by using vocabulary tree using k-means clustering, hierarchy is created. Descriptive vectors of the tree are used for unsupervised learning. Instead, k shows the final number of clusters or quantized cells, it shows the number of children of each node. Before running the k-means algorithm [15], value of k is determined. Training data are divided into k groups. Each group consisted of descriptive vectors, especially closer to the cluster center than the others. This process is done recursively. Each section is divided into k parts, and the tree can be built step by step. As long as the tree is reached the maximum number of levels (l), the process continues. The divisions are done according to the distributing of descriptive vectors that are belong to the parent node. This process is shown in Figure 5.

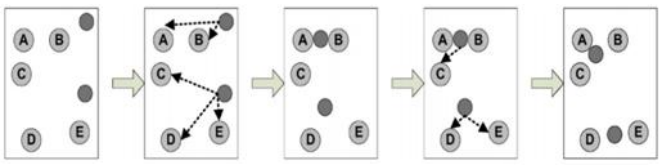

Figure 5. Illustrating process of building a vocabulary tree using k-means (k = 2)

3.2. Setting the TFIDF Weights of the Tree

Once the vocabulary tree is defined, we require to determine a weight w_i to each node i in the vocabulary tree. Here, TFIDF, the product of term frequency and inverse document frequency, is used to assign the weights in the vocabulary tree. The tf-idf weighting algorithm diminishes weight nodes, which appear often in the database as:

$$w_i = tf * \log \frac{N}{N_i} \quad (1)$$

Where N is the total number of images in the database, Ni is the number of images in the database with at least one key-point path through node i, and tf the frequency of occurrence of node i in place of Ni. We define query (qi = niwi) and database (di = miwi) vectors, where ni and mi are the number of key-points vectors of the query and database image, respectively, with a path through node i. After assigning the weights, scoring scheme is defined as:

$$Score\ (q.d) = \left\| \frac{q}{|q|} - \frac{d}{|d|} \right\| \quad (2)$$

Where |.| is L1-norm.

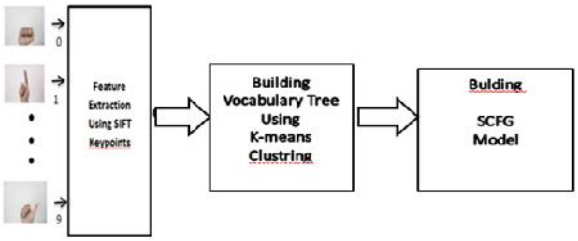

Figure 6. Training stage

In the training phase, we use 100 training images for each hand gesture in the cluttered background. The first test we performed on our database to compute the recognition accuracy rate of our method. Our method has the best recognition results as they show in Table 1. Recognition accuracy rates of 96.27 and recognition time were about 15 ms (mili seconds).

Table 1. Evaluate Performance In f the proposed method on our image database

Posture name	Recognition Accuracy	Recognition Time
One	94.12%	14.1 ms
Two	96.17%	14.3 ms
Tree	95.34%	16.2 ms
Four	97.10%	15.1 ms
Five	95.89%	15.3 ms
Six	94.48%	14.2 ms
Seven	98.15%	16.5 ms
Eight	97.18%	16.4 ms
Nine	97.85%	15.3 ms
Ten	96.80%	14.1 ms

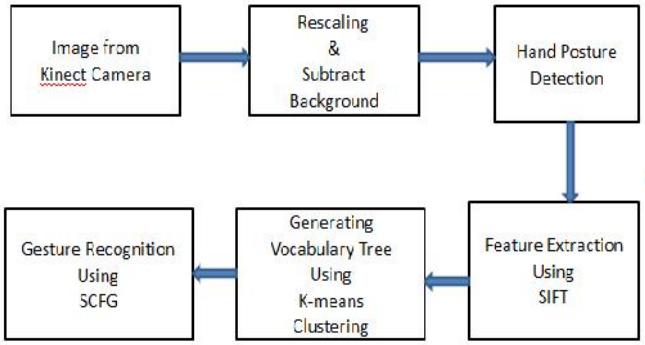

Figure 7. Test sorage

3.3. Stochastic Context-Free Grammars

As a hand gesture is an action composed of a sequence of hand postures, it makes the syntactic approach appropriate to describe the hand gesture in terms of its constituent hand postures. With the syntactic approach, hand gestures can be specified as building up by a group of hand postures in various ways of composition, just as phrases are built up by words. The rules governing the composition of hand postures into different hand gestures can be specified by a grammar. The syntactic approach provides the capability to describe a set of complex hand gestures by using a small group of simple hand postures and a set of grammatical rules.

Stochastic context-free grammars (SCFG) are used in our system to describe the structural and dynamic information about hand gestures. The SCFG is an extension of the context-free grammar (CFG). The difference between SCFGs and CFGs is that for each production rule in SCFGs, there is a probability associated with it. Each SCFG is a four tuple: GS = (VN, VT, PS, and S) where VN and VT are finite sets of non-terminals and terminals; S € VN is the start symbol; PS is a finite set of stochastic production rules each of which is of the form:

$$x \xrightarrow{p} \text{a} \tag{3}$$

Where X € V_N, λ€ V and P is the probability associated with this production rule. The probability P can also be expressed as P(X λ), and it satisfies:

$$\sum_{\text{j}} \text{p}(\text{x} \rightarrow \text{«}j) = 1 \tag{4}$$

Where μj are all of the strings that are derived from X. In SCFG, the notion of context-free essentially means that the production rules are conditionally independent [16]. If a string y € L(GS) is unambiguous and has a derivation with production rules r1, r2, . . . , rk 2 PS, then the probability of y with respect to GS is:

$$p(y \mid G_s) = \prod_{i=1}^{K} p(r_i) \tag{5}$$

If y is ambiguous and has l different derivation trees with corresponding probabilities P1 (y|GS), P2 (y|GS), . . . , Pl (y|GS), then the probability of y with respect to GS is given by:

$$p(y \mid G_s) = \sum_{1}^{l} p_i(y \mid G_s) \tag{6}$$

SCFGs extend CFGs in the same way that Hidden Markov models (HMMs) extend regular grammars. The relation between SCFGs and HMMs is very similar to that between CFGs and non-probabilistic Finite State Machines (FSMs), where CFGs relax some of the structural limitations imposed by FSMs, and because of this, SCFGs have more flexibilities than HMMs [17].

4. Experimental Results

Based on the classified postures from the last section, a simple SCFG is defined to describe the local finger motions, which generate five different gestures: “take”, “unhand”, “Raise”, “drop down”, “Shake”. Each gesture is composed of two postures as illustrated in Figure 8. The SCFG that generates these gestures is defined by $G_G = (V_{NG}, V_{TG}, P_G, S)$, where $V_{NG} = \{S\}$, $V_{TG} = \{P, F, T, V, L, R\}$

And P_G:

$r_{1:S} \xrightarrow{30\%}$ fl , $r_{2:S} \xrightarrow{25\%}$ fp , $r_{3:S} \xrightarrow{15\%}$ fv

$r_{4:S} \xrightarrow{20\%}$ ft , $r_{5:S} \xrightarrow{10\%}$ fr

The terminals P, F, T, V, L, and R stand for the six postures: “palm”, “fist”, “two fingers”, “victory, little finger, “four finger”, and two modes give a gesture. In Figure 8, you can see constituent states of each gesture.

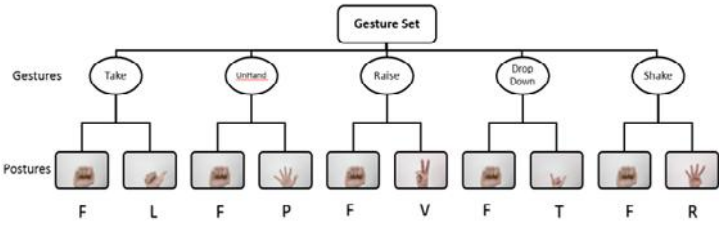

Figure 8. A set of combination of posture that forms gestures

After a string “x” is converted from the postures, we can decide the most likely product rule that can generate this string by computing the probability:

$$p(r \Rightarrow x) = D(Z_r, x)P(r) \quad (7)$$

D (Zr, x) is the similarity between the input string “x” and “z_r”, which is the string derived by the production rule “r”. D (Zr, x) can be computed according to:

$$D(Z_r, x) = \frac{count(Z_r \bigcap x)}{count(Z_r) + count(x)} \quad (8)$$

Table 2. input string "FL"

	D(Zr,x)	$P(r_i)$ i=1,2,…,5	D(Zr,x)* $P(r_i)$
r_1 → fl	(2/6)	30 %	(2/6)*30%=10 %
r_2 → fl	(1/6)	25 %	(1/6)*25%=4.16 %
r_3 → fl	(1/6)	15 %	(1/6)*15%=2.5 %
r_4 → fl	(1/6)	20 %	(1/5)*20%=3.33 %
r_5 → fl	(1/6)	10 %	(1/6)*10%=1.66 %

As shown in Table 2, According to the greatest probability, the gesture represented by this string should be classified as a “Take” gesture.

The flexibility of the SCFG allows the user to change the grammar easily so that other gestures with different combinations of detected postures or more complex gestures can be described. The assignment of the probability to each production rule can also be used to control the “wanted” gestures and the “unwanted” gestures. Greater values of probability could be assigned to the “wanted” gestures. As you can see in Table 3, we do not have any TR movements in a set of defined gestures. However, due to the higher probability we assign it to a desired posture.

Table 3. Input string "TR"

	D(Zr,x)	$P(r_i)$ i=1,2,…,5	D(Zr,x)* $P(r_i)$
$r_1 \rightarrow$ tr	(0/6)	30 %	(0/6)*30%=0 %
$r_2 \rightarrow$ tr	(0/6)	25 %	(0/6)*25%=0 %
$r_3 \rightarrow$ tr	(0/6)	15 %	(0/6)*15%=0 %
$r_4 \rightarrow$ tr	(1/6)	20 %	(1/6)*20%=3.33 %
$r_5 \rightarrow$ tr	(1/6)	10 %	(1/6)*10%=1.66 %

Given that the input string is not equal to any of our defined gestures, but according to the most probable result, we decided to take the gesture shown by this string, is belong to “drop down” gesture.

Finally, for comparison with other methods of gesture recognition, we had a number of articles that we used real-time performance. In the Table 4 we can see, the method is better than other methods for recognition accuracy.

Table 4. Comprising our method with other real time methods

Method	Frame resolution	Background	Recognition time	Recognition accuracy
[18]	320*240	Clutter	0.09-0.11	93.8 %
[[2]]	640*480	Clutter	0.017	96.23 %
[19]	320*240	Wall	0.03	90.0 %
Our method	640*480	Clutter	0.024	96.27 %

5. Conclusion

Provide We have provided a method for recognizing hand gestures using a Microsoft Kinect sensor, the sensor uses infrared rays, and gives information about the depth and of the gray level of rescaled image. According to the image of the depth, Hand can be detected by the use of distinct levels of gray level. We recognize the hand identification process, without impacting of the background or lighting changes on our operations. After detection of hand, we train a vocabulary tree for recognizing Hand gestures. We based on experience and conducted experiments obtained the best parameters (number of levels and branching factor) for vocabulary tree for high accuracy rate of recognition of hand. Then we have proposed context-free grammars for the high level of recognition of Hand gestures.

References

[1] RH Liang, MOM Ouhyoung. *A real-time continuous gesture recognition system for sign language.* Proc. Third IEEE Int. Conf. Autom. Face Gesture Recognit. 1998.

[2] NH Dardas, ND Georganas. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Trans. Instrum. Meas. 60.* 2011: 3592–3607.

[3] H Guan, RS Feris, M Turk. *The Isometric Self-Organizing Map for 3D hand pose estimation*, In: FGR 2006 Proc. 7th Int. Conf. Autom. Face Gesture Recognit. 2006: 263-268.

[4] B Stenger, PRS Mendonca, R Cipolla. *Model-based 3D tracking of an articulated hand.* Proc. 2001 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition. CVPR 2001. 2001; 2.
[5] WDW Du, HLH Li. *Vision based gesture recognition system with single camera.* WCC 2000 - ICSP 2000. 2000 5th Int. Conf. Signal Process. Proceedings. 16th World Comput. Congr. 2000. 2000; 2.
[6] J Segen, S Kumar. Human-computer interaction using gesture recognition and 3D hand tracking. In: Int. Conf. Image Process. 1998: 188-192.
[7] Y Liu, Y Jia. *A robust hand tracking and gesture recognition method for wearable visual interfaces and its applications.* In: Proc. - Third Int. Conf. Image Graph. 2004: 472–475.
[8] D Kim, S Lee, J Paik. Active Shape Model-Based Gait Recognition Using Infrared Images, Int. J. Signal Process. *Image Process. Pattern Recognit.* 2009; 2: 1-12.
[9] B Stenger. Template-Based Hand Pose Recognition Using Multiple Cues. In: PJ Narayanan, S Nayar, HY Shum. *Editors*. Comput. Vis. – ACCV 2006 SE - 55. Springer Berlin Heidelberg. 2006: 551-560.
[10] C Wang, K Wang. Hand Posture Recognition Using Adaboost with SIFT for Human Robot Interaction. In: Lect. Notes Control Inf. Sci. 2008: 317-329.
[11] DG Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis. 2*004; 60: 91-110.
[12] L Juan, O Gwun. A comparison of sift, pca-sift and surf. *Int. J. Image Process.* 2009; 3: 143-152.
[13] YKY Ke, R Sukthankar. *PCA-SIFT: a more distinctive representation for local image descriptors.* Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, 2004. 2004; 2.
[14] M Sonka, V Hlavac, R Boyle. Image processing, analysis, and machine vision. Second edition. Int. Thomson. 1999.
[15] DJC Mackay. Information Theory, Inference, and Learning Algorithms. *Learning.* 2003; 22: 348-349.
[16] A Stolcke. Bayesian Learning of Probabilistic Language Models. 1994.
[17] YA Ivanov, AF Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* 2000; 22: 852-872.
[18] Y Fang, K Wang, J Cheng, H Lu. *A Real-Time Hand Gesture Recognition Method*. Multimed. Expo, 2007 IEEE Int. Conf. 2007: 995-998.
[19] ALC Barczak, F Dadgostar. Real-time hand tracking using a set of cooperative classifiers based on Haar-like features 1 Introduction 2 Invariant Features. *Res. Lett. Inf. Math. Sci.* 2005; 7: 29-42.