# Comparative Study of Various Classifier Models for Gesture Detection & Recognition: A Survey

Prince Dubey
Department of CSE
Technocrats Institute of Technology
Bhopal

Dr. Amit Sinhal
Department of CSE
Technocrats Institute of Technology
Bhopal

**Abstract**
This paper depicts comparison among various human action recognition techniques. Image recognition plays essential role in the field of computer science and language technology. Gesture recognition is the sub branch of image recognition under the human action recognition. Gesture recognition is defined as 'it's a mathematical explanation by a machine". This field have various application like suspicious activity recognition at railway station or airport, facial gesture recognition, gesture based mobile application etc. Gesture recognition can be 2D OR 3D when it is applied to image we can say it is 2d and when applied to video or motion is 3D.the purpose of this paper is to illustrate methods based on human action recognition ,comparative study of action recognition approaches and result of various model in terms of percentage efficiency.

## 1. Introduction

As it is known Human activity [1] recognition is a most important area of computer vision research now these days. The main goal of human activity recognition [2] is to automatically analyze enduring activities from an unknown video which is generally a sequence of image frames. In a simple case where a video is segmented in order to contain only one execution of a human activity, the main aim of the system is to classify the video into its activity category correctly. In most general case, the continuous human activities recognition [3] must be performed by detecting starting and ending times of all occurring activities from an input video. The capability in order to recognize complex human activities from the given videos that enables the construction of many important applications. The Automated surveillance systems which are placed public places such as airports and subway stations etc. usually require detection of abnormal and suspicious activities, which opposed to normal activities. For example, an airport surveillance system must be able to recognize automatically the suspicious activities like "a person is leaving a bag" or "a person is placing something his/her bag in a trash bin." Recognition of human activities also enables the real-time monitoring of Patients, children, and even elderly persons. The erection of gesture-based human computer interfaces and vision-based intelligent environments becomes possible with an activity recognition system also. There are several types of human activities [4] Depending on their complexities, it would be abstractly classify the human activities into four different levels: interactions, gestures, actions, and group activities. Gestures are basic movements of a person's body part, and are basically the atomic components relating to the meaningful motion of a person. "Stretching an arm" and "raising a leg" may be excellent examples of gestures. Actions are mainly single-person activities which may composed of multiple gestures that are organized temporally, for example "walking," "waving," and "punching." Interactions are human activities that involve two or more persons and/or objects. For example, the two persons are fighting is an interaction between two humans and "a person snatching a suitcase from another" is a human-object interaction involving two humans and one object. Finally, syndicate activities are the activities performed by conceptual groups composed of many members or objects: "A group of persons marching," "a group is having a meeting," and "two groups fighting" are typical examples.

The main objective of this paper is to provide a complete summary of state-of-the-art human activities recognition methodologies. As it has been discussed various types of approaches designed for the recognition of different levels of activities. The foregoing review written by Aggarwal and Cai

covered numerous essential low-level components in order to understand the human motion [5], such as tracking and body gesture study.

However, the motion study methods themselves were quite insufficient to exemplify and explain human activities with complicated structures, and most of techniques in 1990s focused on the recognition of motion and simple actions. In this new review, we concentrate on upper level activity recognition techniques designed for the study of human actions, interactions, and bunch of activities, discussing recent research trends in activity recognition.

Figure 1 depict an overview of the tree-structured taxonomy that our review follows. We have chosen an approach-based taxonomy. All activity recognition methods are first organize into two categories: single-layered technique and Hierarchical technique. Single-layered techniques are those that represent and recognize human activities directly based on sequences of images. Due to their character, single-layered approaches are suitable for recognition of gestures.

### 1.1 Single-Layered Approaches

Single-layered approaches detect human activities directly from video data. This type of approaches consider an action as a specific class of image sequences, and recognize the activity from an unknown image sequence (i.e., an input) by label it into its class. Various representation methodologies and matching algorithms have been developed to allow the recognition system to make a correct decision as to whether an image sequence belongs to a specific activity class or not. For recognition from regular videos, most single-layered method have adopted a sliding windows method that group all probable subsequences. Single-layered approaches are most fruitful when a particular sequential pattern [6] that describes an action can be captured from training sequences.

The video volumes are assemble by concatenating image frames along a time axis, and are compared in order to measure their analogy. On the other hand, sequential [7] approaches treat a human action as a sequence of certain observations. More specifically, they represent a human activity as a sequence of feature vectors obtain from images and they identify activities by enquiring for such a sequence. We discuss [8] in Section 2 and compare sequential approaches. Due to their qualities, the main aim of the single-layered method are to examine relatively simple (and short) continuous motion of humans, like walking, jumping, and waving. In this paper, we group [9] single-layered approaches into [10] two classes: space-time technique and sequential technique. Space-time method represents a human activity as a certain 3-D volume in a space-time dimension or a set of characteristics obtain from the volume.

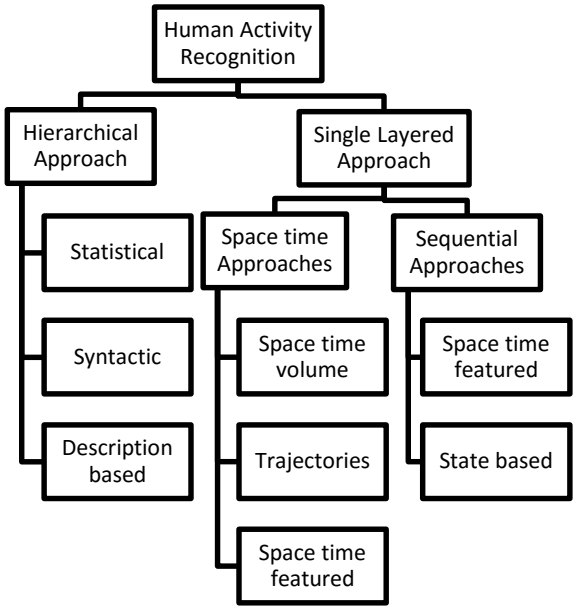



Figure 1. Overview of the tree-structured taxonomy [13]

## 2 Hierarchical Approaches

The main objective behind Hierarchical approaches is to allow the recognition of high-level activities based on the recognition outcome of other basic activities. The determination is to permit the simpler sub activities (also called subevents) that can be modeled comparatively easy to be recognized first, and then to use them for the recognizing upper-level activities. For example, an upper-level interaction like "fighting" may be recognized by detecting a series of many punching and kicking actions. Therefore, in Hierarchical approaches a high-level human activity (e.g., fighting) that the system aims to identify is illustrated in terms of its sub-events (e.g., punching), which may themselves be analyzable until atomicity is acquire. That is, subevents suffice as observations generated by an upper-level activity. The paradigm of hierarchical interpretation not only makes the

recognition process computationally manageable and conceptually understandable; but also reduces repetition in the identification process by reusing recognized subevents several times.

In public, common activity patterns [10] of motion that occur repeatedly during upper level human actions are framed as atomic-level (or primitive-level) actions, and upper-level activities are depicted and identified by concatenating them hierarchically. In most of hierarchical [11] techniques, these atomic actions are identified by acquiring single-layered recognition methods which we represent in the foregoing section. For example, the gestures "stretching hand" and "withdrawing hand" occur usually in human activity, implying that they can turn into good atomic actions for depicting human activities specific as shaking hands or punching. Single-layered approaches equivalent as [12] sequential approaches using HMMs can be safely follow for identification of those gestures.

The primary benefit of Hierarchical method over single-layered method is their capability to identify upper-level activities with more complicated structures. Hierarchical approaches are particularly appropriate for a semantic-level analysis of interplay between humans and/or objects as well as complex group activities. This benefit is a result of two abilities of hierarchical techniques the ability to manage with less training data and the capability to combine earlier knowledge into the representation.

### 2.1 Statistical Approaches

Statistical method use statistical state-based models to recognize activities. In the case of hierarchical statistical method, several layers of state-based models (usually two layers) such as HMMs and DBNs are used to identify action with sequential structures. At the lower layer, atomic activities are identify from chain of feature vectors, just as in single-layered sequential method. As an outcome, a series of feature vectors is converted to a series of atomic actions. The second-level models consider this series of atomic actions, as observations create by the second level models. For each model, a probability of making a series of observations (i.e., atomic-level actions) is calculated to measure the conceivability between the activity and the input image Sequence. Either the Maximum a Posteriori Probability (MAP) or the Maximum Likelihood Estimation (MLE) classifier is built as a result.

### 2.2 Syntactic Approaches

Syntactic method model human action as a string of symbols, where each symbol correlate to an atomic-level action. Homogeneous to the case of Hierarchical statistical method, Syntactic approaches also entail atomic-level actions to be identify first, using any of the foregoing techniques. Human actions are represented as a set of production rules creating a string of atomic actions, and they are recognized by acquiring parsing techniques from the arena of Programming languages. Stochastic Context-free grammars (SCFGs) and Context-free grammars (CFGs) have been used by foregoing researchers to identify high-level activities. The production principle of CFGs naturally lead to a Hierarchical representation and identification of the activities.

### 2.3 Description-Based Approaches

A description-based approach is a recognition method that explicitly preserve human activities' Spatio-temporal structures. It represents upper-level human activity in terms of simpler activities that produce the activity (i.e., sub events), describing their temporal, spatial, and logical relationships. That is, description-based technique frame a human activity as a happening of its sub event that satisfies definite relations. Therefore, identification of the activity is performed by searching the sub events fulfilling the relations stated in its representation. All description-based approaches are naturally hierarchical (since they use sub events to represent human activities), and they are capable to handle activities with simultaneous structures.

## 3. Convolutional Neural Networks

A 2D convolution is simply the application [9] of a mask to a 2D image, conceptually the same as a "blur" operation in computer graphics. The prototype of a Sequoia task which implements a 2D convolution is as follows.

$$c[m][n]=\sum_{u}\sum_{v}(A[m+u][n+v].H[u][v])$$

A 2D convolution problem, producing an output image of size M*N, can be decomposed into a set of

parallel 2D convolution sub problems, each computing a non-overlapping region of the output image of size S*T. This is the decomposition that is used to implement this algorithm in Sequoia.

### 3.1. 3D Convolution

In 2D CNNs, convolutions [9] are applied on the 2D character maps to figure out properties from the spatial dimensions only. Method applied for video learning, it is necessary to catch the motion data encoded in any adjacent frames. For this, we introduce 3D convolutions in the convolution stages of CNNs to tally properties from both spatial and temporal dimensions. The 3DC is obtain by convolving a 3D kernel to the cube formed by bundle many adjacent frames together. By this arrangement, the feature maps in the convolution layer are connected to many adjacent frames in the precedent layer, thereby securing motion information.

## 4. Literature Survey

**Di Wu et al** proposed and implement an "efficient Gesture Segmentation Model using Dynamic Neural Networks" [13]. This paper describes a novel method called Deep Dynamic Neural Networks (DDNN) for multimodal gesture recognition A hierarchical dynamic model based on a Hidden Markov Model (HMM) is implemented for concurrent gesture segmentation and identification. In which input is taken in the form of skeleton joint information, depth and RGB images. Different from most common approaches that rely on the structure of complex handcrafted characteristics, the method learns uppers-level Spatio-temporal description using deep neural networks appropriate for the input

**Natadlia Never ova et** al implemented an efficient Gesture Detection and Recognition technique using Multi-Scale Approach [14]. Here in this paper propose a generalized approach to human gesture recognition based on multiple data modalities such as depth video, articulated pose and speech. In our system, each gesture is decomposed into large-scale body motion and local subtle movements such as hand articulation. The idea of learning at multiple scales is also applied to the temporal dimension, such that a gesture is considered as a set of characteristic motion impulses, or dynamic poses. Each modality is first processed separately in short spatial-temporal blocks, where discriminative data-specific features are either manually extracted or learned. Finally, we employ a Recurrent Neural Network for modeling large-scale temporal dependencies, data fusion and ultimately gesture classification.

| | **Skeleton** | **RGB-D** | **Fusion** |
|---|---|---|---|
| 3 set skeletal & HOG, Boosted classifier [1] | 0.791 | - | 0.822 |
| 3D skeletal pose & HOG, MRF[19] | 0.79 | - | 0.827 |
| Dense trajectory (HOG,HOF,MBH) [21] | - | 0.792 | - |
| Template base, Random Forest Classifier[15] | - | - | 0.747 |
| Fisher Vector, Dynamic Programming[22] | 0.745 | - | - |
| Max Margin (MMHCRF)[5] | - | 0.9251 | - |
| Hidden Conditional Random Field (HCRF) [5] | - | 0.8760 | - |
| Independent Subspace Study, RF [18] | - | 0.649 | - |
| PHOG, SVM, HMM [16] | 0.454 | 0.462 | 0.597 |
| Deep Neural Networks[17] | 0.747 | 0.637 | 0.804 |
| DDN[13] | 0.779 | 0.717 | 0.809 |

Table 1 Results of various Techniques [13]

**Necati Cihan et al** proposed Template Based Random Forest Classifier for the Recognition and Detection of Gestures [15]. This paper presents a framework for spotting and recognizing continuous human gestures. Skeleton based features are extracted from normalized human body coordinates to represent gestures. These features are then used to construct spatio-temporal template based Random Decision Forest models. Finally, predictions from different models are fused at score level to improve overall recognition performance. Our method has shown competitive results on the CHALEARN 2014 Looking at People dataset. Trained on a dataset of 20 gesture vocabulary and 7754 gesture samples, our method achieved a Jaccard Index score of 74:6% on the test set, reaching 7th place among contenders. Among methods that exclusively used skeleton based features, our method obtained the highest recognition performance.

**Bin Liang et al** implemented and proposed Skeletal Joint and Motion Trial Model for the Multi-model

Gesture Recognition [16]. This paper proposes a novel approach to multi-modal gesture recognition by using skeletal joints and motion trail model. The approach includes two modules, i.e. spotting and recognition. In the spotting module, a continuous gesture sequence is segmented into individual gesture intervals based on hand joint positions within a sliding window. In the recognition module, three models are combined to classify each gesture interval into one gesture category. For skeletal model, Hidden Markov Models (HMM) and Support Vector Machines (SVM) are adopted for classifying skeleton features. For depth maps and user masks, we employ 2D Motion Trail Model (2DMTM) for gesture representation to capture motion region information. SVM is then used to classify Pyramid Histograms of Oriented Gradient (PHOG) features from 2DMTM. These three models are complementary to each other. Finally, a fusion scheme incorporates the probability weights of each classifier for gesture recognition.

**Camille Monnier et.al** proposed an efficient Multi-Scale Boosted Detector for the efficient Robust Gesture Recognition [17].Here an approach to detecting and recognizing gestures in a stream of multi-modal data. Our approach combines a sliding window gesture detector with features drawn from skeleton data, color imagery, and depth data produced by a first generation Kinect sensor. The detector consists of a set of one-versus-all boosted classifiers, each tuned to a specific gesture. Features are extracted at multiple temporal scales, and include descriptive statistics of normalized skeleton joint positions, angles, and velocities, as well as image-based hand descriptors.

**Guang Chen et al** uses and implemented Gesture Detection and Recognition using Unsupervised Randomization and Discrimination [18]. The goal of approach is to identify semantically meaningful contents from dense sampling spatio temporal feature space for gesture recognition. To achieve this, we develop three concepts under the random forest framework: un-supervision; discrimination; and randomization. Un-supervision learns spatio temporal features from four channels (grayscale, depth, gradient, surface normal) of RGB-D video in an unsupervised way. Discrimination extracts the information in dense sampling spatio-temporal space effectively. Randomization explores the dense sampling spatio-temporal feature space efficiently.

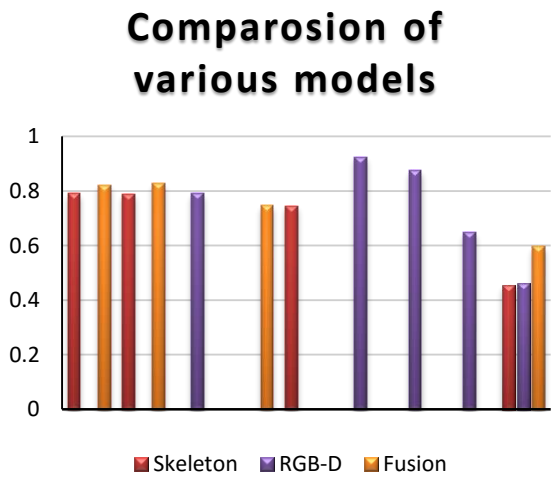
**Ju Yong Chang et al** proposed a new and efficient technique for Recognition and Detection of Gestures using Multi-Model Data [19]. A new gesture recognition method using multimodal data. The approach solves a labeling problem, which means that gesture categories and their temporal ranges are determined at the same time. For that purpose, a generative probabilistic model is formalized and it is constructed by non-parametrically estimating multi modal densities from a training dataset. In addition to the conventional skeletal joint based features, appearance information near the active hand in the RGB image is exploited to capture the detailed motion of fingers. The estimated log-likelihood function is used as the unary term for our Markov random field (MRF) model. The smoothness term is also incorporated to enforce temporal coherence of our model.

**Di Wu et al** also implemented an efficient technique for Gesture Segmentation and Recognition using Dynamic Neural Networks [20]. The purpose of this paper is to describe a novel method called Deep Dynamic Neural Networks (DDNN) for the Track 3 of the Cha learn Looking at People 2014 challenge. A generalized semi supervised hierarchical dynamic framework is proposed for simultaneous gesture segmentation and recognition taking both skeleton and depth images as input modules. First, Deep Belief Networks (DBN) and 3D Convolutional Neural Networks (3DCNN) are adopted for skeletal and depth data accordingly to extract upper level Spatio-temporal features. Then the learned representations are used for estimating emission probabilities of the Hidden Markov Models to infer an action sequence. The framework can be easily extended by including an ergodic state to segment and recognize video sequences by a frame-to-frame mechanism, rendering it possible for online segmentation and recognition for diverse input modules. Some normalization details pertaining to preprocessing raw features are also discussed.

**Georgios Evangelidis** et al uses and implemented Continuous Gesture Recognition from articulated process [22]. This paper addresses the problem of

continuous gesture recognition from articulated poses. Unlike the common isolated recognition scenario, the gesture boundaries are here unknown, and one has to solve two problems: segmentation and recognition. This is cast into a labeling framework, namely every site (frame) must be assigned a label (gesture ID). The inherent constraint for a piece-wise constant labeling is satisfied by solving a global optimization problem with a smoothness term. For efficiency reasons, we suggest a dynamic programming (DP) solver that seeks the optimal path in a recursive manner.

**Xiaojiang Peng et al** proposed Super Vector Representation based technique for the Recognition and Detection of Gestures [2]. This paper focuses on describing our method designed for both track 2 and track 3 at Looking at People (LAP) challenging. We propose an action and gesture spotting system, which is mainly composed of three steps: (i) temporal segmentation, (ii) clip classification, and (iii) post processing. For track 2, we resort to a simple sliding window method to divide each video sequence into clips, while for track 3, we design a segmentation method based on the motion study of human hands. Then, for each clip, we choose a kind of super vector representation with dense features. Based on this representation, we train a linear SVM to conduct action and gesture recognition.




Graph1: accuracy based on different model

## 5. Conclusion

Human activities recognition plays significant role in the field of computer vision explore. This survey paper depict all the technique that are implemented for the Detection and Recognition .This paper work analyze various techniques of gesture recognition and also Comparative study of their performance. This work motivated to go for further research. Here data domain is images and videos as data set. Base work uses CNN and Chalearn data set, this data set contains videos of various human gestures. Due to wide area of research on image recognition number of method available for research. Table1 depict the comparative study of some of these methods. These methods shows that Neural Network have strong recognition capability ,CNN is Dynamic Neural network also have higher pattern recognition rate, by enhancing feature extraction methods lead to very efficient results this feature of recognition motivates to go for further future Research.

## References


[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions From Movies," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008..

[2] J. Liu, J. Luo, and M. Shah, "Recognizing Realistic Actions From Videos," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1996-2003, 2009.

[3] Y. Wang and G. Mori, "Max-Margin Hidden Conditional Random Fields For Human Action Recognition," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 872-879, 2009.

[4] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, "Automatic Annotation of Human Actions In Video," Proc. 12th IEEE Int'l Conf. Computer Vision, pp. 1491-1498, 2009.

[5] Y. Wang and G. Mori, "Hidden Part Models For Human Action Recognition: Probabilistic Versus Max Margin," IEEE Trans. Pattern Study and Machine Intelligence, Volume. 33, no. 7, pp. 1310-1323, July 2011.

[6] Y. LeCun, F. Huang, and L. Bottou, "Learning Methods For Generic Object Recognition With Invariance To Pose And Lighting," Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, 2004

[7] J. Kittler, M. Hatef, R.P. Duin, and J. Matas, "On Combining Classifiers," IEEE Trans. Pattern Study and Machine Intelligence, Volume. 20, no. 3, pp. 226-239, Mar. 1998.

[8] S. Lazebnik, C. Achmid, and J. Ponce, "Beyond Bags Of Features: Spatial Pyramid Matching For Recognizing Natural Scene Categories," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2169-2178, 2006.

[9] Shuiwang Ji, Wei Xu, Ming Yang," 3d Convolutional Neural Networks For Human Action Recognition", IEEE transactions on pattern stud and machine intelligence, Volume. 35, no. 1, pp. 221-231,January 2013.

[10] L.K. Hansen and P. Salamon, "Neural Network Ensembles," IEEE Trans. Pattern Study And Machine Intelligence, Volume. 12, no. 10, pp. 993-1001, Oct. 1990.

[11] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Object Recognition With Cortex-Like Mechanisms," IEEE Trans. Pattern Study and Machine Intelligence, Volume. 29, no. 3, pp. 411-426, Mar. 2007.

[12] K. Schindler and L. Van Gool, "Action Snippets: How Many Frames Does Human Action Recognition Require?" Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

[13] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao," Deep Dynamic Neural Networks For Multimodal Gesture Segmentation And Recognition", IEEE Transactions on Pattern Study and Machine Intelligence, Volume 38, No. 8, August 2016.

[14] Natalia Neverova, Christian Wolf, Giulio Paci, Giacomo Sommavilla, Graham W. Taylor, Florian Nebout," A Multi-Scale Approach To Gesture Detection And Recognition", ICCV, 2013.

[15] N. C. Camg¨oz, A. A. Kindiroglu, and L. Akarun, "Gesture Recognition Using Template Based Random Forest Classifiers," in Proc. Comput. Vis. Workshops, 2014, pp. 579–594

[16] B. Liang and L. Zheng, "Multi-Modal Gesture Recognition Using Skeletal Joints And Motion Trail Model," in Proc. Comput. Vis. Workshops, pp. 623–638,2014.

[17] C. Monnier, S. German, and A. Ost, "A Multi-Scale Boosted Detector For Efficient And Robust Gesture Recognition," in Proc. Eur. Conf. Comput. Vis. Pattern Recog. Workshops, pp. 491–502,2014.

[18] G. Chen, D. Clarke, M. Giuliani, A. Gaschler, D. Wu, D. Weikersdorfer, and A. Knoll, "Multi-Modality Gesture Detection And Recognition With Un-Supervision, Randomization And Discrimination," in Proc. Comput. Vis. Workshops, pp. 608–622,2014.

[19] J. Y. Chang, "Nonparametric Gesture Labeling From Multi-Modal Data," in Proc. Eur. Conf. Comput. Vision Pattern Recog. Workshops, pp. 503–517,2014.

[20] D. Wu and L. Shao, "Deep Dynamic Neural Networks For Gesture Segmentation And Recognition," in Proc. Eur. Conf. Comput. Vis. Pattern Recog. Workshops, 2014, pp. 552–571.

[21] X. Peng, L. Wang, and Z. Cai, "Action And Gesture Temporal Spotting With Super Vector Representation," in European Conference on Computer Vision and Pattern Recognition Workshops, 2014.

[22] G. D. Evangelidis, G. Singh, and R. Horaud, "Continuous Gesture Recognition From Articulated Poses," in Computer Vision-ECCV 2014 Workshops. Springer,pp. 595–607.2014.