LOCALIZATION AND RECOGNITION OF DYNAMIC HAND GESTURES BASED ON HIERARCHY OF MANIFOLD CLASSIFIERS

M. Favorskaya*, A. Nosov, A. Popov

Institute of Informatics and Telecommunications, Siberian State Aerospace University, 31 Krasnoyarsky Rabochy av., Krasnoyarsk, 660014 Russian Federation - favorskaya@sibsau.ru, alexander@nosov.org, vm_popov@sibsau.ru

Commission WG V/5, WG III/3

ABSTRACT:

Generally, the dynamic hand gestures are captured in continuous video sequences, and a gesture recognition system ought to extract the robust features automatically. This task involves the highly challenging spatio-temporal variations of dynamic hand gestures. The proposed method is based on two-level manifold classifiers including the trajectory classifiers in any time instants and the posture classifiers of sub-gestures in selected time instants. The trajectory classifiers contain skin detector, normalized skeleton representation of one or two hands, and motion history representing by motion vectors normalized through predetermined directions (8 and 16 in our case). Each dynamic gesture is separated into a set of sub-gestures in order to predict a trajectory and remove those samples of gestures, which do not satisfy to current trajectory. The posture classifiers involve the normalized skeleton representation of palm and fingers and relative finger positions using fingertips. The min-max criterion is used for trajectory recognition, and the decision tree technique was applied for posture recognition of sub-gestures. For experiments, a dataset "Multi-modal Gesture Recognition Challenge 2013: Dataset and Results" including 393 dynamic hand-gestures was chosen. The proposed method yielded 84–91% recognition accuracy, in average, for restricted set of dynamic gestures.

1. INTRODUCTION

During three decades after appearance of graphical user interface with mouse and keyboard, the sensor and display technologies evolve persistent improving and expanding the novel devices ranging from very large displays for design and educational projects to small smartphones or smart watches for conventional consuming. All these devices keep pushing researchers to develop new interaction techniques based on natural human possibilities, first of all, using human gestures and body movements as more natural and intuitive communication between people and devices. Hand gestures have been studied for a long time since 1990s starting with indispensable attributes such as different color gloves in order to provide the simplified tracking of hands and fingers in 3D environment.

Design of formal visual languages with the goal of easy human–computer communication through the use of graphics, drawings, or icons meets the challenge of complicated technical implementation because of variety of locations, shapes, overlaps of hands and cluttered background in a scene. Human gestures are classified as the head, the hand, and the body gestures. Each group assumes special capturing, tracking, and recognition methods. The upper-body gestures are represented in two forms: as natural and artificial gestures. The natural gestures are uncertain, with cultural and local diversity while the artificial gestures are more comprehensible for predefined actions. Also existing gesture systems are classified as encumbered (when a user ought to hold an external device to make gestures (Lu et al., 2014)), touch-based (systems with the touch-screen and different commands), and vision-based (allow users to make gestures without contact). The detailed review of vision-based hand gesture recognition is presented by Rautaray and Agrawal (Rautaray and Agrawal, 2015).

Before recognition, a segmentation stage is necessary. Any gesture segmentation is a difficult process, especially for dynamic gestures. The segmentation process is characterized by ambiguities, when the start and end time instants of dynamic gesture are difficult identified in continuous sequence (Yang et al., 2007), and a spatio-temporal variability, which is caused by gesture variety in shape, location, duration, and trajectory, even for the same person at different time instants (Kosmopoulos and Maglogiannis, 2006). However, a simultaneous procedure of gesture segmentation and recognition is discussed in some researches (Kim et al., 2007).

Our contribution deals with building of hierarchy of manifold classifiers including two levels: the trajectory classifiers in any time instants and the posture classifiers in selected time instants. The trajectory classifiers include skin detector, normalized skeleton representation of hand/hands, and motion history representing by motion vectors normalized through predetermined directions (8 and 16 in our case). Each dynamic gesture is separated into a set of sub-gestures in order to predict a trajectory and remove those samples of gestures, which do not satisfy to current trajectory. The posture classifiers involve the normalized skeleton representation of palm and fingers and relative fingers position using fingertips.

The rest of this paper is organized as follows. Section 2 reviews the existing methods of dynamic gesture capturing and tracking as well as methods of gesture recognition. The proposed gesture

---

*   Corresponding author

trajectory classifiers are described in Section 3. Section 4 provides a discussion about the posture classifiers. Experimental results are situated in Section 5. Finally, Section 6 concludes the paper.

2. RELATED WORK

In this section, two mutual issues will be briefly reviewed representing the dynamic gesture capturing and tracking methods and the recognition methods in Sections 2.1–2.2, respectively.

2.1 Methods of Gesture Capturing and Tracking

The Hidden Markov Models (HMMs) and the Dynamic Time Warping (DTW) are the most popular methods for dynamic gesture recognition. Many researchers employed the HMM for gesture capturing in the spatio-temporal framework successfully as well as for gesture recognition. The HMM and its modifications was especially popular in 1990s. The HMM-based threshold model was proposed by Lee and Kim (Lee and Kim, 1999). The start and end points were fixed by a likelihood threshold calculated for predefined gesture models. The DTW is a method for sequence comparison, initially used in various applications. For gestures with varying lengths, the DTW warps the tested trajectory in order to match with a predetermined template of an exemplary gesture trajectory. A precursor to the DTW was the Longest Common Subsequence (LCS) method of alignment. It was successfully applied by Stern et al. (Stern et al., 2010) for TV remote control. The LCS algorithm was developed for matching sub-word sequences in documents temporally, using feature distance costs. The LCS is more robust to noise and outliers the DTW in computational speed. Instead of a complete mapping between all points, the LCS algorithm ignores a point without good matching. The classifier uses the "spotted" gesture trajectories or moving window of trajectory points.

Alon et al. (Alon et al., 2005) developed the method with core of Dynamic Space-Time Warping (DSTW) algorithm. The DSTW algorithm as an extension of the DTW aligns a pair of query and model gestures in both space and time dimensions. A warping path in time is aligned with detection the best hand candidate region in every query frame by dynamic programming. The DSTW algorithm models the multiple candidate feature vectors (hypotheses) for hand location in each frame. The warping path in the spatial and the temporal dimensions has some constraints such as the boundary conditions, the temporal continuity, and the temporal monotonicity. The system worked in cluttered background, multiple moving objects, and multiple skin-colored image regions. However, the starting and ending frame of each gesture were pointed manually.

Krishnan et al. (Krishnan et al., 2010) proposed to form a gesture spotting network using the individual gesture models and the adaptive threshold model learnt from Adaptive Boosting algorithm. This technique was evaluated for hand gestures spotting from continuous accelerometer data streams and recognizing by the HMM based on the adaptive threshold model with precision of 78% and recall of 93%.

The attempts of simultaneous gesture spotting and recognition attract some researchers due to a good idea to reduce an unavoidable time delay between segmentation and recognition for on-line continuous gesture recognition. The forward scheme computing a Competitive Differential Observation Probability (CDOP) between a gesture and a non-gesture was proposed by Kim et al. (Kim et al., 2007). This scheme modelled a sequentially variant time-series of gestures using the HMM and Self Organizing Maps (SOMs). The authors computed observation probability of gesture or non-gesture by use a number of continuing observations within the sliding window for several observations. This helps to avoid an undesirable effect of an abrupt change of observations within a short interval. An association mapping technique was used to estimate a correlation measure between 2D shape data and 3D articulation data. Then the gestures were recognized by the trained HMM.

The Conditional Random Fields (CRFs) as an alternative approach to the HMM are studied by Elmezain et al. (Elmezain et al., 2010) for simultaneous spotting and recognition of digits. Such interpretation can be useful for dynamic gesture recognition. The CRFs are undirected graphical models for labeling sequential data, which overcome the weakness of maximum entropy Markov models. The authors conducted good idea about initially constructed the non-gesture pattern by the CRFs due to the CRFs uses a single model for the joint probability of the sequences. As a result, they obtained good reliability accuracy estimations of 96.51% and 90.49% for recognition of isolated and meaningful gestures, respectively.

Extraction and tracing of hand region using entropy analysis was utilized by Lee et al. (Lee et al., 2004). It was one of the first attempts for proposition of video-based non-contact interaction techniques. A hand region was extracted based on color information and entropy distribution. In common case, a gesture tracking can be based on various techniques, e.g., fingertips tracking for restricted set of gestures, hand's position tracking by optical flow method, motion energy analysis to estimate a dominant motion of hand, and adaptive skin color model and Motion History Image (MHI), splitting a feature trajectory into sub-trajectories.

2.2 Methods of Dynamic Gesture Recognition

In general, two types of recognition techniques, 2D shape data and 3D articulation data, are possible for dynamic gesture recognition. Bobick and Davis (Bobick and Davis, 2001) recognized human movements by 2D temporal templates including a motion energy image and the MHI. Hu moments were descriptors of the temporal templates. A human gesture recognition based on fitted quadratic curves as a place of centre points of skin region and 2D foreground silhouettes was proposed by Dong et al. (Dong et al., 2006). The principal disadvantage consists in the evident dependence of the obtained 20 features from a viewing angle between a human and cameras. Also a gesture may be classified in a conventional manner, when a sequence of postures is processed, using principal component analysis, and then it is recognized applying Finite State Machines (FSMs) also called as a finite-state automata (Hong et al., 2000).

Bhuyan (Bhuyan, 2012) proposed a concept of key frames detection, in which a hand changes its shape significantly. Next keyframes are extracted by measuring a shape similarity using a Hausdorff distance measure. Thus, such key frame selection eliminates redundant frames. Each FSM constructed through the training stage corresponds to a particular gesture. Recognition is executed by matching only a current keyframe state with the states of different FSMs obtained during the training stage. All

the FSMs in the gesture vocabulary use an angular radial transformation shape descriptor that speeds up the recognition in many times.

Dynamic Bayesian Network (DBN) is a generalized class of probabilistic models including the HMM and Kalman filter or particle filter as special cases. Suk et al. (Suk et al., 2010) proposed the DBN model for recognition of hand gestures implementing a control of media players and slide presentations. Ten isolated one-hand and two-hand gestures are enough simple, and the authors received the recognition rate of 99.59%. More practical continuous gesture recognition was addressed based on a cyclic spotting network connecting with gesture DBN. A Viterbi dynamic programming method was used to recognize gesture and detect the start and end points of gesture simultaneously.

Al-Rousan et al. (Al-Rousan et al., 2010) studied the dynamic gestures of Arabic sign language using two-stage scheme, when, first, a group of gestures is recognized and, second, a gesture within a group is recognized. A set of spatial features was extracted including a hand region, coordinates of its centre, a direction angle of hand region, and a hand vector representing a shape of hand. The authors considered these features invariant to scale and transition. The HMM was used for hand recognition. The authors divided all features into two types: six simple features and seventeen complex (in vector representation) features. The recognition rate for the signer-dependent achieved 92.5% while for the signer-independent was 70.5%.

3D modelling of hand gesture is usually connected with multiple cameras shooting during training stage or Kinect application in order to obtain the depth images of a hand. For example, Keskin et al. (Keskin et al., 2011) proposed a 3D skinned mesh model with a hierarchical skeleton, consisting of 19 bones, 15 joints and 21 different parts for representation of American Sign Language. The methodology is based on fitting 3D skeleton to the hand. Random decision forests were trained on animated 3D hand model for pixel classification during the testing stage.

Holte and Moeslund (Holte and Moeslund, 2008) proposed to build and analyze the harmonic shape context as 3D motion primitives, which are received from motion characteristic instances of gestures. The authors used 3D image data from a range camera to achieve invariance to viewpoint. For this purpose, a family of orthogonal basis functions in a form of spherical harmonics was applied. 3D motion primitives representing as the strings of letters and/or digits with different lengths were compared by probabilistic edit distance method. An overall recognition rate of 82.9% was achieved under invariance of 3D location.

3. GESTURE TRAJECTORY CLASSIFIERS

Introduce assumptions, according to which a video sequence contains a hand gesture:

1. A hand gesture is performed on approximately uniform background.
2. A distance between a camera and a hand is nearly constant so that scale factor is non-significant.
3. Consider that a moving hand appears in video sequence, if a motion is continuous in a predetermined interval (1–2 s).
4. A hand gesture is performed in a priori known region of frame.
5. A moving hand is a dominant moving object.
6. A duration of moving is longer that L_1 frames but not more than L_2 frames with the temporal continuity and monotonicity.
7. All types of gestures are a priori known.
8. Images of hand gestures are captured with a single video camera and then processed by a single computer.

Dynamic gesture spotting is very difficult task without simplifying guesses, e.g., when the gesture trajectories are initiated by a button press or a long interval without motion. Li and Greenspan (Li and Greenspan, 2011) had solved the endpoint localization of dynamic gestures using a multi-scale gesture model representing as 3D spatio-temporal surfaces.

A motion history of dynamic gesture is a set of trajectory classifiers of sub-gestures, which model a realistic gesture by elements from special vocabulary. First, hand localization ought to be implemented (Section 3.1). Second, skeleton representation of gesture is built (Section 3.2). Third, trajectory classifiers are constructed (Section 3.3).

3.1 Hand Localization

For hand localization, it is reasonable to use two classifiers based on skin and motion detection. The skin-like color regions may be detected in a scene using various color spaces such as Red Green Blue (RGB), YUV (Y is a luminance component (brightness), U and V are the blue and red difference samples, respectively), YCbCr (Cb and Cr are the blue and red chromatic components, respectively), Hue Saturation Value (HSV), normalized RGB, and Log opponent (uses the base 10 logarithm to convert RGB values) color spaces (Favorskaya, 2013). The YCrCb color space was chosen for skin classifier. The linear equations, which determine the boundaries of skin color classifier, are provided by Equation 1:

$$
\begin{cases}
Cr \geq -2(Cb+24) \\
Cr \geq -4(Cb+32) \\
Cr \geq -(Cb+17) \\
Cr \geq 25(Cb+Q_1) \\
Cr \geq Q_3
\end{cases}
;
\begin{cases}
Cr \leq \dfrac{220-Cb}{6} \\
Cr \leq \dfrac{4}{3}(Q_2 - Cb) \\
Cr \geq 0.5(Q_4 - Cb)
\end{cases}
\tag{1}
$$

where Q_1–Q_4 = additional parameters computed by Equations 2–3 in dependence of Y value:

$$
\begin{cases}
\text{if } Y > 128 \text{ then } Q_1 = -2 + \dfrac{256-Y}{16} \, , \\
Q_2 = -20 - \dfrac{256-Y}{16}, \; Q_3 = 6, \; Q_4 = -8
\end{cases}
\tag{2}
$$

$$
\begin{cases}
\text{if } Y \leq 128 \text{ then } Q_1 = 6 \; Q_2 = 12 \, , \\
Q_3 = 2 + Y/32, \; Q_4 = -16 + Y/16
\end{cases}
\tag{3}
$$

For reliable detection of hand/hands in frames, some authors proposed the approach based on a prior facial skin analysis as an individual skin tuning (Li et al., 2013). Such initial procedure is very useful for hand localization because the articulated user can be a single end-user of gesture recognition system, and this procedure can be considered as an adaptive

step. Face detection based on Adaboost algorithm is fast and accuracy method proposed by Viola and Jones (Viola and Jones, 2001). As in any boosting algorithm, a cascade classifier is adopted during 2–3 iterations in order to detect the most possible regions of human faces. Additionally, eyes detection can be recommended in regions similar to skin color. Then hands are detected in enlarged surrounding region, which sizes are chosen empirically. After hand detection using skin classifiers, the morphological processing is recommended for improvement "broken" skin regions or skin regions with "holes" (Favorskaya and Nosov, 2014).

The use of Microsoft Kinect camera is in the area of interest for many segmentation tasks. The simple background subtraction can be applied under assumption of enough number of frames in static scene. First, a background accumulation is executed based on maximum values of depths in a set of frames. As a result, a background model can be constructed. Notice, an accurate background model is not required for hand capturing that means a possibility of on-line background model building. Second, the moving objects are extracted using a background subtraction model. Third, the object boundaries are computed based on detection and merger of contour components in order to receive the closed boundaries of moving objects. The depth map cannot be used only for body parts contour extraction but also for body silhouette building. The last possibility is useful for the body gesture recognition. For hand segmentation, the color distribution is analysed into segmented regions that increases the accuracy of palm and fingers localization in a cluttered background.

3.2 Skeleton Representation of Gesture

In this research, the idea of compact and informative description of dynamic gesture is conducted. For this purpose, a skeleton representation of hand including elbow, wrist, palm, and fingers is built. The normalized skeleton representation is invariant to shape but do not invariant to position in 3D space. However, in skeleton representation a "central" point as a centre point of circle, which is inscribed in a palm image region, can be defined in the most cases very fast.

A binary hand gesture image is an object with multi-linked polygonal shape. In this case, a skeleton representation is built using a term "maximum empty circle" (Mekhedov and Mestetskiy, 2010). For polygonal shape F, a maximum empty circle is any circle B, which is fully inscribed into a shape F such that other circle B' inscribing into a shape F does not include a circle B. Thus, a skeleton of polygonal shape F is a set of centres of maximum empty circles. A radial function $R(x, y)$ defining a radius value in any skeleton point (x, y) is determined on a skeleton of an object.

Geometrically, a skeleton of polygonal shape is a graph including nodes (points in the XOY plane) and edges (lines joining some nodes pairs). A straight line or a parabolic arc will be the edges of such graph. A degree of any edge is equal 1, 2, or 3. The existing effective algorithms build a skeleton by time O (n log n), where n is a number of nodes in a polygon shape. A skeleton building of 2D shape is described in details by Mekhedov and Mestetskiy (Mekhedov and Mestetskiy, 2010). An execution time of a skeleton building depends directly from a number of nodes in a polygon binary shape. The procedure of shape approximation based on Douglas-Peucker algorithm is represented and implemented in our previous research (Favorskaya and Nosov, 2014) with good speed/accuracy

numerical results. After skeleton building, an additional procedure called pruning is implemented in order to remove noisy lines. Then a central point in palm image can be easily defined.

3.3 Trajectory Classifiers Based on Motion History

Let DG = {DG$_1$, DG$_2$, ..., DG$_n$} be a set of available dynamic gestures represented by skeleton vectors. Each dynamic gesture DG is composed from sub-gestures differing by directions and acceleration values from other sub-gestures. A set of sub-gestures generates a vocabulary, which helps to describe a realistic dynamic gesture by its model representation. A sequence of sub-gesture forms a gesture model. Due to a hand image is a compact set of mass points, a trajectory can be described by a single unique point, which can be a centroid of mass of a palm or a central point in skeleton representation (Section 3.2). A sub-gesture description SG is a set of vectors SG = {$\mathbf{SG}_1$, $\mathbf{SG}_2$, ..., $\mathbf{SG}_m$}. The beginning of each following vector $\mathbf{SG}_{j+1}$ is the ending of the previous vector $\mathbf{SG}_j$; thus, a connected set of directions performs a total direction of a sub-gesture. Each vector $\mathbf{SG}_j$ includes two components: a length L_j and an angle α_j (between vector direction and the axis OX) in relative coordinates (Equation 4):

$$L_j = \sqrt{\left(x_{j+1} - x_j\right)^2 + \left(y_{j+1} - y_j\right)^2}; \quad \alpha_i = \arctan \frac{y_{j+1} - y_j}{x_{j+1} - x_j} \qquad (4)$$

where x_j, y_j = coordinates of current branch point
  x_{j+1}, y_{j+1} = coordinates of following branch point

Then a sub-gesture SG is recalculated by compression and normalization procedures in order to provide invariance to affine transform. The compression procedure rejects non-essential vectors with small value L_j according to a predetermined threshold value. The normalization procedure provides a normalization of vectors lengths (total sum of vectors lengths is equal 1) by Equation 5:

$$L_j^{nr} = L_j \bigg/ \sum_{j=1}^{n} L_j \qquad (5)$$

and a normalization of vectors directions, when a current vector direction is replaced by one of normalized directions from a unit vectors set. Such unit vector $\mathbf{U}_j$ has a view of Equation 6:

$$\mathbf{U}_j = \left\{1, \alpha_j^{nr}\right\}; \quad \alpha_j^{nr} = j \cdot 2\pi / Z \qquad (6)$$

where Z = a number of directions ($Z = 8$ or $Z = 16$)

Let the temporal representation of all gesture samples be $\mathbf{G}_{c,k}$, $c = 1, 2, ...,$ where C is a number of temporal classes of hand gestures and $k = 1, 2, ..., K$ is a number of samples into a temporal class. The reference temporal template TT_c estimates dissimilarity $D(\cdot)$ by min-max criterion in the temporal dimension by Equation 7:

$$TT_c = \arg \min_k \arg \max_m D\left(\mathbf{G}_{c,k}, \mathbf{G}_{c,m}\right) \qquad (7)$$

Template-based classification technique is announced as follows. The extracted and processed temporal representation is compared with reference temporal template for each class c

(Equation 7). Afterward, distances to all classes are computed. If a minimal distance is below a predetermined threshold value, then a gesture belongs to some class, otherwise it is concerned to a non-gesture class.

During the testing stage, the extracted sub-gestures help to predict a type of gesture. Several candidates may be determined. The posture classifiers of palm and fingers provide the final decision.

4. GESTURE POSTURE CLASSIFIERS

Each sub-gesture ought to include the detailed information about palm and fingers position. Consider that this analysis will be accomplished at the end of sub-gesture. Such information clarifies the class of dynamic hand gesture. Posture classifiers are discussed in Section 4.1. A procedure of rule-based recognition is situated in Section 4.2. A set of rules is applied to select the best candidate.

4.1 Posture Classifiers

The posture classifiers are constructed based on skeleton representation of hand. As one can see from Figure 1, the fingertips can be detected as endpoints. Then the relative locations of skeleton segments of each finger are analyzed, in particular lengths of fingers $S = \{s_1, s_2, s_3, s_4, s_5\}$ and angles $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ between the neighbouring fingers. A length value and a view of skeleton line of finger determine a position of finger. A set of angles may be restricted to one or two values. These values are graduated by threshold: if $TH_\theta < 3°$, then the fingers are grouped. Notice that not all gestures may have full description, some fingers may be invisible.
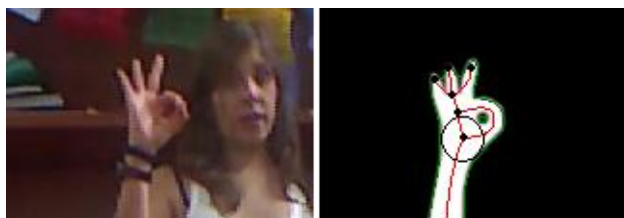

Figure 1. Original image (left), image with keypoints (right)

Figure 1 illustrates a skeleton building based on a binary image. Use of gray-scale image of hand permits to detect a skeleton for fingers overlapping a palm. Therefore, a posture classifier includes relative location of a centre point of a palm (a centre of inscribed circle), coordinates of fingertips, based on which it is possible to determine the fingers stretched, half-bend, half-closed, or closed and group, separate, cross, or loop. In the same manner, two hands can be analyzed, if a gesture is presented using both hands.

4.2 Rule-based Recognition

In spite of variety of 3D hand position, a hand model includes a fixed number of components, and each finger configuration is associated with a finger pose. Each hand posture is described by a rule capturing a hand configuration and involving the following parameters with available values:

1. The coordinates of centre point of a palm.
2. A set of lengths $S = \{s_1, s_2, s_3, s_4, s_5\}$. If $(s_i > TH_s)$ and $(s_i$ is a straight line), then a finger is stretched. If $(s_i > TH_s)$

and $(s_i$ is a curve), then a finger is half-bend or half-closed. If $(s_i > TH_s)$ and $(s_i,\ _{sj}$ are closed curves), then two fingers form a loop. If $(s_i \le TH_s)$ and $(s_i$ is a straight line), then a finger is closed.
3. A set of angles $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. If $\theta_i > TH_\theta$, then a two fingers are separated. If $(\theta_i \le TH_\theta)$ and (a thickness of two fingers is less than the double thicknesses of one finger), then two fingers are crossed.
4. Relative coordinates of visible fingertips.
5. Detection a finger group based on a convexity analysis of a hand contour. For example, one can detect the groups from two, three, or four closed fingers estimating a thickness of group.

Items 1–5 mentioned above describe each sub-gesture as a set of finger orientations in the XOY plane (up, down, towards, or side), a set of finger inter-relations (grouped, separated, looped, or crossed), and a set of basic finger poses (stretched, half-bend, half-closed, or closed). The main positions of fingers based on depth performances and depicted in Table 1 were proposed by Mo and Neumann (Mo and Neumann, 2006).

Category	Position			
Basic fingers orientation	Up	Down	Towards	Side
Basic finger inter-relations	Group	Separate	Cross	Loop
Basic finger poses	Half-bend	Bend	Half-closed	Closed

Table 1. Main positions of fingers

As a result, a sequence of sub-gesture descriptors can be received. If dominant singularity/singularities of a gesture are determined during the training stage, then the testing stage is simplified. The decision trees based on the proposed rules are used for recognition. A hierarchy description of sub-gestures including trajectory analysis finds a good mapping in decision trees. Due to rejection of non-suitable branches, the recognition is a fast procedure. When a dynamic gesture is finished, the descriptions of sub-gestures are finally analyzed in the temporal dimension, and the final decision is concluded. In spite of difficult analysis, a continuous sequence provides high data volume that objectively leads to better recognition results.

5. EXPERIMENTAL RESULTS

For experiments, a part of the dataset "Multi-modal Gesture Recognition Challenge 2013: Dataset and Results"[1] containing dynamic gestures was used. This large video dataset includes 13, 858 gestures from a lexicon of 20 Italian gesture categories recorded with a Kinect[TM] camera, providing the audio, skeletal model, user mask, RGB and depth images (Escalera et al., 2013). For better visibility in Figures 2–9, fragments of images with sizes 140×105 pixels were cut from the test video files.

---

[1] http://gesture.chalearn.org/2013-multi-modal-challenge

Figure 2. Sample00827, frames 1175-1231

Figure 6. Sample00804, frames 216-251

Figure 3. Sample00824, frames 164-211

Figure 7. Sample00802, frames 913-970

Figure 4. Sample00825, frames 189-232

Figure 8. Sample00816, frames 1368-1408

Figure 5. Sample00831, frames 921-1015

Figure 9. Sample00803, frames 432-475

The software tool "DynGesture", v. 1.52 was developed using C# language in environment "Visual Studio 2012". The software includes two main algorithms. The segmentation algorithm contains a hand/hands localization based on skin classifiers, a topological skeleton building, a skeleton representation of palm and fingers with keypoints, a building of trajectory classifiers, and a vector description of sub-gestures. The recognition algorithm calculates the posture classifiers based on five proposed rules (Section 4.2) providing a recognition of main positions of fingers at the end of each sub-gesture. Then a sequence of sub-gesture descriptors is analyzed by decision tree procedure. Samples from each category involving 20–40 dynamic gestures were divided into the training and the testing sets in relation 20% and 80%, respectively. The results of gesture segmentation and gesture recognition are located in Tables 2 and 3, respectively. These estimations include True Recognition (TR), False Rejection Rate (FRR), False Acceptance Rate (FAR), and time costs of algorithms as average estimators for processing of a single frame. Experiments were executed using computer with the processor Intel(R) Core(TM) i5 750 2.67 GHz. The TR, FRR, and FAR estimators were received based on frames segmented manually and recognized by expert.

Video file	TR (%)	FRR (%)	FAR (%)	Time (ms)
Sample00827	92.9	3.3	8.5	9.2
Sample00824	86.7	3.1	6.7	8.6
Sample00825	93.2	7.7	9.1	10.4
Sample00831	90.4	2.3	4.6	7.7
Sample00804	93.5	3.9	6.4	8.0
Sample00802	83.3	6.6	10.2	9.3
Sample00816	84.3	8.3	4.5	4.2
Sample00803	90.8	5.2	9.1	8.5

Table 2. Gesture segmentation results and temporal estimators

As one can see from Table 2, all true and false segmentation results have close values in ranges of 84-93% and 3-10%, respectively. The deviation is explained by some blurred frames in video files, for example, "Sample00802" and "Sample00816". The temporal values of frame processing have the same order.

Video file	TR (%)	FRR (%)	FAR (%)	Time (ms)
Sample00827	84.9	8.3	8.9	4.4
Sample00824	78.6	7.1	5.2	5.6
Sample00825	91.1	8.8	6.7	5.7
Sample00831	81.6	5.3	4.5	4.2
Sample00804	83.6	7.5	9.1	4.8
Sample00802	93.4	5.8	5.1	5.3
Sample00816	85.4	7.3	4.6	4.0
Sample00803	89.4	8.7	5.2	6.1

Table 3. Gesture recognition results and temporal estimators

Gesture recognition results representing in Table 3 demonstrate the dependence from more simple or complex trajectory of dynamic gesture and visible shape of a hand. In spite of all difficulties, the final results are promising due to the analysis of a sequence of sub-gesture descriptors, which represent a single dynamic gesture. The temporal estimators show a possibility of real-time implementation of the designed software tool.

6. CONCLUSION

The proposed method for localization and recognition of dynamic gestures is based on two-level classifiers including the trajectory classifiers in any time instants and the posture classifiers for sub-gesture extraction in selected time instants. Our efforts were directed on building of such classifiers, which are invariant to scale, rotation, and shift in the XOY plane. Skeleton representation of wrist, palm, and fingers provides a description of special points (e.g., fingertips or centre point of a palm), based on which some rules were formulated. Motion history of trajectory and a set of posture descriptions of sub-gestures are the input information for decision trees. A dataset "Multi-modal Gesture Recognition Challenge 2013: Dataset and Results" was used for experiments. The numerical results were obtained for 393 dynamic gestures applicable in learning systems of sign languages as well as in human-computer interaction systems. The proposed approach yielded 84–91% recognition accuracy for restricted set of dynamic gestures.

REFERENCES

Alon, J., Athitsos, V., Yuan, Q., Sclaroff, S., 2005. Simultaneous Localization and Recognition of Dynamic Hand Gestures. *7th IEEE Workshops on Application of Computer Vision*, 2, pp. 254-260.

Al-Rousan, M., Al-Jarrah, O., Al-Hammouri, M., 2010. Recognition of dynamic gestures in arabic sign language using two stages hierarchical scheme. *International Journal of Knowledge-based and Intelligent Engineering* 14(3), pp. 139-152.

Bhuyan, M.K., 2012. FSM-based recognition of dynamic hand gestures via gesture summarization using key video object planes. *World Academy of Science, Engineering and Technology* 6(8), pp. 724-735.

Bobick, A., Davis, J., 2001. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* 23 (3), pp. 257-267.

Dong, Q., Wu, Y., Hu, Z., 2006. Gesture recognition using quadratic curves. In: *The 7th Asian Conference on Computer Vision*, pp. 817-825.

Elmezain, M., Al-Hamadi, A., Michaelis, B., 2010. A Robust Method for Hand Gesture Segmentation and Recognition Using Forward Spotting Scheme in Conditional Random Fields. In: *The 10th IEEE International Conference on Pattern Recognition*, pp. 3850-3853.

Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., Athitsos, V., Escalante, H.J., (2013) Multi-modal Gesture Recognition Challenge 2013: Dataset and Results. In: *15th ACM International Conference on Multimodal Interaction*, pp. 445-452.

Favorskaya, M., 2013. Visual system of sign alphabet learning for poorly-hearing children. In: Kountchev, R., Iantovics, B. (eds) *Advances in Intelligent Analysis of Medical Data and Decision Support Systems*, SCI 473, pp. 23-39.

Favorskaya, M., Nosov, A., 2014. Hand-gesture Description Based on Skeleton Representation and Hu Moments. *Frontiers in Artificial Intelligence and Applications* 262, pp. 431-440.

Holte, M.B., Moeslund, T.B., 2008. View invariant gesture recognition using 3D motion primitives. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 797-800.

Hong, P., Turk, M., Huang, T.S., 2000. Gesture Modeling and Recognition Using Finite State Machines. In: *The IEEE Conference on Face and Gesture Recognition*, pp. 410-415.

Keskin, C., Kira, F., Kara, Y.E., Akarun, L. 2011. Real time hand pose estimation using depth sensors. In: *The IEEE International Conference on Computer Vision Workshops*, pp. 1228-1234.

Kim, D., Song, J., Kim D., 2007. Simultaneous gesture segmentation and recognition based on forward spotting accumulative HMMs. *Pattern Recognition* 40(11), pp. 3012-3026.

Kosmopoulos, D.I., Maglogiannis, I.G., 2006. Extraction of Mid-Level Semantics from Gesture Videos using a Bayesian Network. *International Journal of Intelligent Systems Technologies and Applications* 1(3-4), pp. 359-375.

Krishnanm, N.C., Lade, P., Panchanathan, S., 2010. Activity Gesture Spotting Using a Threshold Model Based on Adaptive Boosting. In: *The IEEE International Conference on Multimedia and Expo*, pp. 155-160.

Lee, H., Kim, J., 1999. An HMM-based threshold model approach for gesture recognition, *IEEE Trans. Pattern Anal. Mach. Intell*. 21(10), pp. 961-973.

Lee, J.S., Lee, Y.J., Lee, E.H., Hong, S.H., 2004. Hand region extraction and Gesture recognition from video stream with complex background through entropy analysis. *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1513-1516.

Li, H., Greenspan, M., 2011. Model-based segmentation and recognition of dynamic gestures in continuous video streams. Pattern Recognition 44 (8), pp. 1614-1628.

Li, J., Zheng, L., Chen, Y., Zhang, Y., Lu, P., 2013. A Real Time Hand Gesture Recognition System Based on the Prior Facial Knowledge and SVM. *Journal of Convergence Information Technology* 8(11), pp. 185-193.

Lu, Z., Chen, X., Li, Q., Zhang, X., Zhou, P., 2014. A Hand Gesture Recognition Framework and Wearable Gesture-Based Interaction Prototype for Mobile Devices. *IEEE Transactions on Human-Machine Systems* 44(2), pp. 293-299.

Mekhedov, I., Mestetskiy, L., 2010. Skeleton of a Multi-ribbon Surface. In: Taniar, D., Gervasi, O., Murgante, B., Pardede, E., Apduhan, B.O. (eds) *Computational Science and Its Applications*, LNCS 6016, pp. 557-573.

Mo, Z., Neumann, U., 2006. Real-time Hand Pose Recognition Using Low-Resolution Depth Images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2, pp. 1499-1505.

Rautaray, S.S., Agrawal, A., 2015. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review* 43(1), pp 1-54.

Stern, H., Frolova, D., Berman, S., 2010. Hand Gesture Recognition for TV Remote Control using Tree-Based Ensemble and LCS Classifiers. *International Conference on Image Processing, Computer Vision and Pattern recognition*, pp. 687-693.

Suk, H.I., Sin, B.K., Lee, S.W., 2010. Hand gesture recognition based on dynamic Bayesian network framework. *Pattern Recognition* 43(9), pp. 3059-3072.

Viola, P., Jones, M., 2001. Rapid Object Detection using a Boosted Cascade of Simple Features [C]. In: *The IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.511-518.

Yang, H.D., Park, A.Y., Lee, S.W., 2007. Gesture Spotting and Recognition for Human–Robot Interaction. *IEEE Trans. on Robotics* 23(2), pp. 256-270.