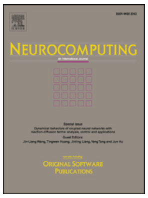
Neurocomputing 267 (2017) 311–319

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



A four dukkha state-space model for hand tracking

Kian Ming Lim[a,*], Alan W.C. Tan[b], Shing Chiang Tan[a]

[a] Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, Melaka 75450, Malaysia
[b] Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh Lama, Melaka 75450, Malaysia

ARTICLE INFO

Article history:
Received 23 June 2015
Revised 19 January 2017
Accepted 4 June 2017
Available online 9 June 2017

Communicated by Zhang Zhaoxiang

Keywords:
Hand tracking
Four dukkha
State-space model

ABSTRACT

In this paper, we propose a hand tracking method which was inspired by the notion of the four dukkha: birth, aging, sickness and death (BASD) in Buddhism. Based on this philosophy, we formalize the hand tracking problem in the BASD framework, and apply it to hand track hand gestures in isolated sign language videos. The proposed BASD method is a novel nature-inspired computational intelligence method which is able to handle complex real-world tracking problem. The proposed BASD framework operates in a manner similar to a standard state-space model, but maintains multiple hypotheses and integrates hypothesis update and propagation mechanisms that resemble the effect of BASD. The survival of the hypothesis relies upon the strength, aging and sickness of existing hypotheses, and new hypotheses are birthed by the fittest pairs of parent hypotheses. These properties resolve the sample impoverishment problem of the particle filter. The estimated hand trajectories show promising results for the American sign language.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Hand tracking is one of the challenging tasks in computer vision that aims to estimate the continuous hand motion in hand gesture video. Hand tracking is important in many applications, including human-computer interaction, human behaviour analysis and hand gesture recognition. In this work, hand tracking in sign language recognition is considered.

Sign language is a visual communication means used by hearing impaired community to communicate. In real world, there are limited hearing people who are able to communicate in sign language. In view of this, researchers have been developing sign recognition systems to bridge the communication gap between the hearing and hearing impaired communities. Generally, sign language recognition can be categorized into isolated sign recognition and continuous sign recognition. A sign comprises hand motion and hand shapes (manual components), as well as facial expressions, head motion and body postures (non-manual components). Prior to recognizing the hand gesture, a sign language recognition system must be able to locate the hands. In this work, we focus on hand tracking of isolated sign gestures using only the manual components.

Hand movement is an important cue in sign language recognition. In the context of an automated sign language recognition system, a sign may span several image frames within a gesture video. For that reason, a state-space model hand tracking method inspired by the birth, aging, sickness and death (BASD) life cycle is proposed and depicted in Fig. 1. Unlike most other state-space model, the proposed BASD state-space hand tracking method maintains multiple hypotheses whereby each hypothesis models the hand location, velocity and age. In the BASD model, each hypothesis goes through the process of birth, aging, sickness and death. The mean of top tier of the fittest surviving hypothesis is the estimated target hand location. The novelties of the proposed method are:

- A novel nature-inspired computational intelligence method to address complex real-world tracking problem.
- Hypotheses which are unfit (due to aging or sickness) undergo culling where hypotheses with low fitness score are eliminated.
- New hypothesis will be established by surviving (parent) hypotheses to overcome the sample impoverishment problem.

The paper is organized as follows: Initially, a review of state-of-the-art hand tracking methods in isolated sign language recognition is provided in Section 2. Subsequently, the details of the proposed BASD hand tracker are described in Section 3. Experiments and discussions are then reported in Section 4. In the same section, the database used in the experiments and the performance evaluation are discussed. Finally, conclusion is drawn in Section 5.

[*] Corresponding author.
E-mail addresses: kmlim@mmu.edu.my (K.M. Lim), wctan@mmu.edu.my (A.W.C. Tan), sctan@mmu.edu.my (S.C. Tan).

http://dx.doi.org/10.1016/j.neucom.2017.06.012
0925-2312/© 2017 Elsevier B.V. All rights reserved.

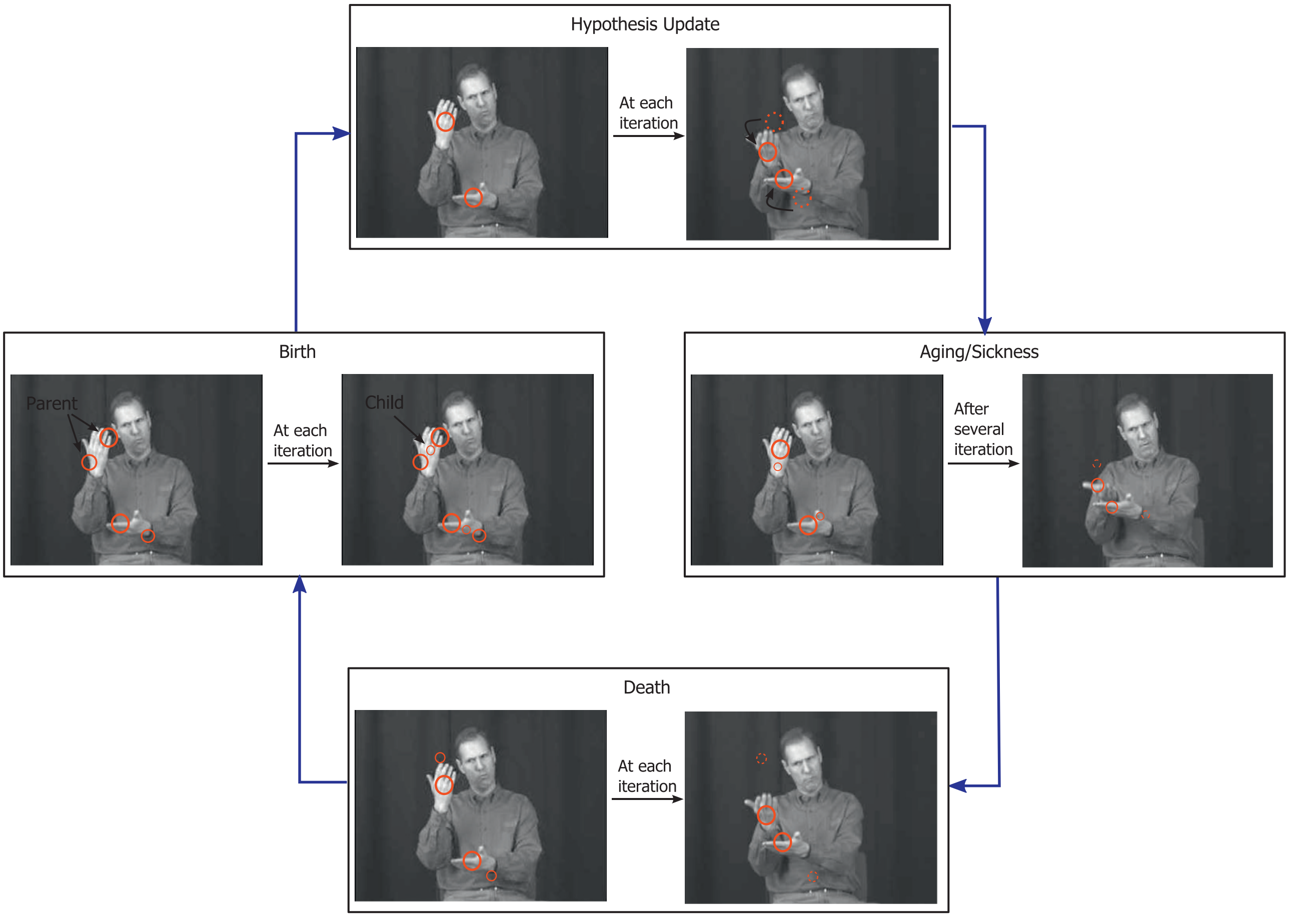


Fig. 1. Overview of proposed BASD hand tracking.

2. Related work

A variety of methods have been researched to extract the features for sign language recognition. Almost all began by detecting the location of hands via hand tracking approaches and many such hand trackers utilize human skin color as it is unique in comparison to the other colors. For example, Chen et al. [1] proposed a fusion of skin color detection, edge detection and motion detection by logical AND operation in their hand tracking algorithm. A skin color distribution is used to segment the hand based on the L*a*b color model [2]. In [3], Zhang and Huang combined skin color and super pixel information to extract the hand region. Even though skin color is easy to differentiate, hand tracking solely based on this feature may fail due to other exposed body parts (e.g. face or the arms) having the same skin color.

Inspired by the success of state-space methods in visual tracking tasks, researchers have begun to apply it to hand tracking. For example, Gaus and Wong [4] employed Kalman Filter to detect hand to head and hand to hand occlusion regions. Park et al. [5] utilized a depth sensor and performed hand tracking using Kalman Filter. Shan et al. [6] adopted the mean shift embedded particle filter as a non-linear posterior density estimator for real time hand tracking. Belgacem et al. [7] likewise embedded optical flow as a penalisation method into particle filter for sign language recognition. Campr et al. [8] used joint particle filter to calculate a combined likelihood model of hands and head. Morshidi and Tjahjadi [9] presented a hand tracking method based on gravity optimized particle filter. The literature demonstrates that particle filter is well-suited to hand tracking applications, given its capability to model non-linear probability distribution, although the performance is highly dependent on suitably chosen dynamic and observation models.

Apart from state-space methods, some other well-known visual tracking methods were also considered for hand tracking. Jang et al. [10] proposed hand tracking in depth images by computing a hand weighted depth probability. They employed continuous adaptive mean shift (CAMSHIFT) algorithm as their tracking procedure. Yoo et al. [11] likewise used CAMSHIFT to track hands in their interactive large-scale display system. Kolsch and Turk [12] introduced KLT feature tracking based hand tracker and the method performed well in unconstrained indoor and outdoor environments. Chen et al. [13] extended the work by using the KLT tracking to efficiently update the search window of an improved CAMSHIFT tracking method. Elsewhere, Chen et al. [14] applied a region growing technique to segment the hand region in depth images and used mean-shift algorithm to track the hand region. Park et al. [15] extracted the candidate hand regions from depth images and chose the best candidate based on the color and shape feature. Then, a boundary tracking method based on Generalized Hough Transform was proposed to track the hand. A recent work by Kishore et al. [16] detected the hand position using optical flow method. Active contour shape features were extracted and input to a fuzzy inference engine for recognition.

Recent researches also focus on color and depth information acquired using RGB-D sensors. Jangyodsuk et al. [17] proposed to utilize Histograms of Oriented Gradients (HOG) for hand shape representation and applied Dynamic Time Warping (DTW) to perform sign language recognition using Kinect sensor. Zhang et al. [18] used Histogram of Oriented Displacement to describe the hand trajectories, and multi-SVM for classification on the sign language recorded using Microsoft Kinect. Likewise, Sun et al. [19] collected an American Sign Language dataset using Microsoft Kinect sensor. In their work, HOG and optic flow features were employed to represent the hand appearance and motion information. Additionally, several features obtained from Kinect such as body pose, hand shape, and hand motion were also utilized. Then, a latent support vector machine model was proposed to classify the signs based on both color image and depth map captured by the sensor. A similar work by Liu et al. [20] employed Microsoft Kinect sensor to collect Chinese Sign Language dataset. Four skeleton joints were used as the input to a long short-term memory architecture for recognition. Despite additional depth information, the RGB-D sensors are less cost effective. For this reason, this work turns to a publicly available dataset recorded using less expensive intensity camera.

3. Proposed BASD hand tracker

Dukkha (suffering) is the first noble truths introduced by Buddha in his first sermon. Birth, aging, sickness and death (BASD) are universal sufferings, as are sorrow, grief, despair, separation and unaccomplished desires. These sufferings are the inevitable cycle of the human life. Some people who come across this teaching may consider it pessimistic. However, Buddhists consider that as neither optimistic nor pessimistic, but realistic. The Buddha's teachings are not just about sufferings; rather, the teachings go on to advise us on how we can eradicate it. Interested readers are referred to Ref. [21] for more information on the subject.

As soon as we are born into the world, we grow, and in the process of growing, we learn from our environment and experiences. However, due to our unique genetic makeup and circumstances, the fitness of our body (or health conditions) are different. Some are strong while others are weak. Some easily become sick while others are healthier and possibly live a longer life. Inspired by these facts, we propose a tracking method to emulate the effect of BASD, and apply it to track the hand gestures in isolated sign language videos.

3.1. Hypothesis update

The proposed BASD hand tracker is a state space model that maintains multiple hypotheses. Each hypothesis is represented as a state vector. Specifically, let the state vector $\mathbf{x}_t^n = [x_t^n \; y_t^n \; u_t^n \; v_t^n \; a_t^n]^T$ be the nth hypothesis in the tth frame containing the location coordinates (x_t^n, y_t^n), and velocities (u_t^n, v_t^n), where u_t^n and v_t^n represent the velocity in the x-direction and y-direction, respectively, and a_t^n is the age associated with the hypothesis. We model the gesture motion as a hypothesis update, given by

$$\begin{aligned}\mathbf{x}_t^n = \begin{bmatrix} x_t^n \\ y_t^n \\ u_t^n \\ v_t^n \\ a_t^n \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1}^n \\ y_{t-1}^n \\ u_{t-1}^n \\ v_{t-1}^n \\ a_{t-1}^n \end{bmatrix} + \begin{bmatrix} n_x \\ n_y \\ n_u \\ n_v \\ 1 \end{bmatrix} \\ &= f(\mathbf{x}_{t-1}) + \mathbf{n}_t, \quad {}_{n=1,2,\ldots,N}^{t=1,2,\ldots,T} \end{aligned} \tag{1}$$

where N is the number of hypothesis, T is the number of frames in the video, $n_x, n_y \sim N(0, \sigma_{xy}^2)$ and $n_u, n_v \sim N(0, \sigma_{uv}^2)$ are the noise terms, and σ_{xy}^2 and σ_{uv}^2 are the noise variance of displacement and velocity, respectively. The age of a new hypothesis, a_t^n, is assigned the value 1, and thereafter incremented by 1 from one frame to the next.

3.2. Hypothesis propagation

To evaluate the fitness of a hypothesis, we define a fitness function. Fitness of human relates to the health condition and strongly correlates with the age. By the same token, the fitness of a hypothesis is related to its age. Let the aging function of a hypothesis be defined as

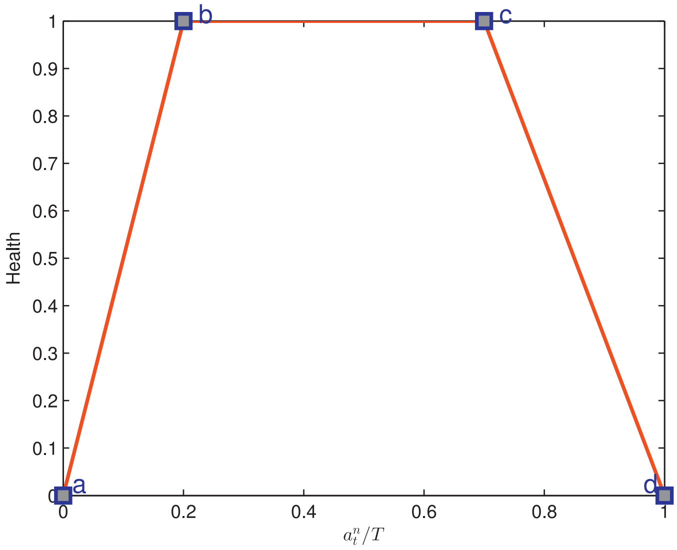


Fig. 2. Sickness function.

$$aging(\mathbf{x}_t^n) = \gamma - \frac{a_t^n}{T} \tag{2}$$

where γ is a constant slightly greater than 1 so that a younger hypothesis is deemed as fitter than an older one. In a human life cycle, the two periods in which one is more prone to sickness are during infancy and old age. Therefore, to evaluate the health of a hypothesis, we model the sickness function as a trapezoidal shaped function of age. The sickness function is defined as

$$sickness(\mathbf{x}_t^n; a, b, c, d) = \max\left(\min\left(\frac{\frac{a_t^n}{T} - a}{b - a}, 1, \frac{d - \frac{a_t^n}{T}}{d - c}\right), 0\right) \tag{3}$$

where a, b, c, d are four scalar parameters that determine the trapezoidal shape. The sickness function is depicted in Fig. 2, where the x-axis represents the value of a_t^n/T and the y-axis represents the fitness of a hypothesis with regards to its age.

In addition to aging and sickness, the fitness of a hypothesis in the proposed BASD hand tracker also considers the strength of the hypothesis. To evaluate the strength of the hypothesis, we make the assumption that, if the hypothesis is located near the hand, then it is strong; otherwise, it is weak. Therefore, hand detection is needed in order to examine whether a hypothesis is located near the hand.

In sign language, when a signer performs a sign gesture, the head of the signer may move even more than the hands, thereby adversely affecting the performance of hand tracking. To overcome this, we apply the Viola and Jones [22] face detection algorithm to locate, and thereafter, remove the face region of the signer prior to further processing.

To detect the hand, we apply the background subtraction method in our previous work [23]. In statistics, the median is the numerical value separating the higher half of the data sample from the lower half. The median, as compared to the mean, is a measure that is especially robust in the presence of outlier in video processing. The mode, on the other hand, is the most common value in the data sample. In the context of background detection, the assumption that moving objects (the hands) are not always located at the same point is made. Thus, a fusion of the median and mode filters for background detection is employed. Specifically, the average of the median and mode of every pixel of the image sequence in modelling the background image is computed. The

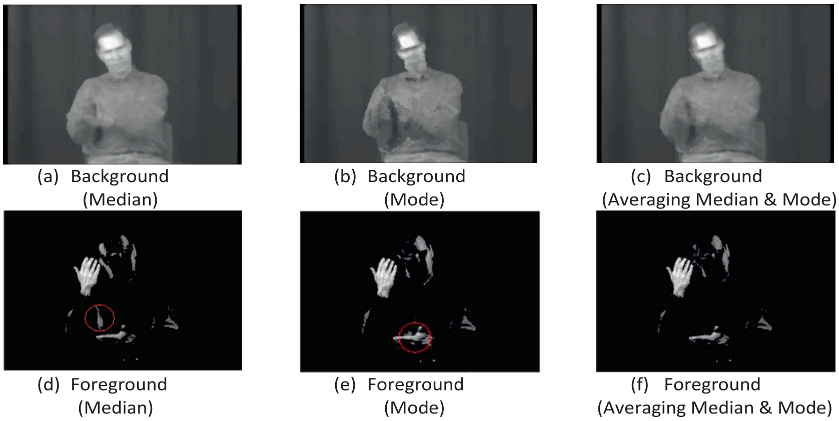

(a) Background (Median)
(b) Background (Mode)
(c) Background (Averaging Median & Mode)
(d) Foreground (Median)
(e) Foreground (Mode)
(f) Foreground (Averaging Median & Mode)

Fig. 3. Comparison of background subtraction using median, mode and average of two.

difference between the background image and the original image above a preset threshold produces the foreground image, which corresponds to regions where the hands are located. In the experiments, it is clear that the combination of these two statistical measures produces the best result, as illustrated in Fig. 3. After the foreground sign language sequences are obtained, the absolute difference of every two subsequent frames is computed to obtain the hand motion region.

After the hand motion region is detected, and based on the estimated location (x_t^n, y_t^n) of the state vector $\mathbf{x}_t^n$, an image patch $\mathcal{P}$ of the size $r \times r$ centered at this location is extracted. The fitness of a hypothesis with regards to its strength is based on the summed and normalized intensity values I of all pixels in $\mathcal{P}$, which is computed as

$$h(\mathbf{x}_t^n) = \frac{\sum_{x,y\in\mathcal{P}} I(x,y)}{r^2} \tag{4}$$

In addition, we also take into account the strength of a hypothesis from the previous iteration, i.e.,

$$strength(\mathbf{x}_t^n) = \alpha \times h(\mathbf{x}_t^n) + (1-\alpha) \times h(\mathbf{x}_{t-1}^n) \tag{5}$$

where $0 < \alpha < 1$ is the weightage parameter. Finally, combining the aging, sickness and strength functions, we define the overall fitness of a hypothesis as

$$\begin{aligned} z_t^n &= fitness(\mathbf{x}_t^n) \\ &= strength(\mathbf{x}_t^n) \times aging(\mathbf{x}_t^n) \times sickness(\mathbf{x}_t^n; a, b, c, d) \end{aligned} \tag{6}$$

In each iteration, $k\%$ of the hypotheses with the lowest fitness score will be eliminated. The birth of a new hypothesis begins with the identification of the candidate parent hypotheses. To that end, we identify the nearest m pairs of the surviving parent hypotheses by computing the pairwise distance as

$$d(\mathbf{x}_t^n, \mathbf{x}_t^{n'}) = \sqrt{(x_t^n - x_t^{n'})^2 + (y_t^n - y_t^{n'})^2}, \qquad n, n' = 1, \ldots, N \tag{7}$$

For each of the m pairs of hypotheses, a number of new hypotheses are birthed at the midpoint of the parents $(\mathbf{x}_t^p, \mathbf{x}_t^q)$, i.e.,

$$Newhypothesis, \mathbf{x}_t^* = \begin{bmatrix} \dfrac{x_t^p + x_t^q}{2} \\ \dfrac{y_t^p + y_t^q}{2} \\ \dfrac{u_t^p + u_t^q}{2} \\ \dfrac{v_t^p + v_t^q}{2} \\ 1 \end{bmatrix} \tag{8}$$

Algorithm 1 BASD Hand Tracker.

```
Initialization: ${(\mathbf{x}_1^n, z_1^n)}_{n=1}^N$
for $t = 2 : T$ do
    for $n = 1 : N$ do
        Hypothesis update: $\mathbf{x}_t^n = f(\mathbf{x}_{t-1}^n)$ (refer to (1))
        Fitness evaluation:
        $z_t^n = strength(\mathbf{x}_t^n) \times aging(\mathbf{x}_t^n) \times sickness(\mathbf{x}_t^n; a, b, c, d)$
        (refer to (6))
    end for
    State estimation: $j\%$ of the fittest hypotheses $\rightarrow \hat{\mathbf{x}}_t$ (refer to (9))
    Death: $k\%$ of hypotheses with lowest fitness score die
    Birth:
    Surviving parent hypotheses $\mathbf{x}_t \rightarrow \mathbf{x}_t \cup \mathbf{x}_t^*$ (refer to (8))
end for
```

At the end of this stage, the total number of hypothesis (parents and children) will be the same as before. Algorithm 1 outlines the proposed BASD hand tracking algorithm.

At the initialization stage of the algorithm, the initial motion of foreground objects are obtained by finding the absolute difference of the first two frames, and Otsu [24] method is employed to binarize the frame difference. The connected components in the binary image are then acquired and sorted in descending order of the area. The centroid with the largest area in the sorted list represents the initial location of the hands. Then, the initial location coordinates $\mathbf{x}_1^n$ of hypothesis n corresponding to the first frame are set at this centroid and equal weights $z_1^n = 1/N$ are assigned to all hypotheses.

In the subsequent frames $(t = 2, 3, \ldots, T)$, all hypotheses step through the algorithm in four main stages, namely hypothesis update, fitness evaluation, estimation and death/birth. The final estimated state $\hat{\mathbf{x}}_t$ in each time step is computed as

$$\hat{\mathbf{x}}_t = \sum\nolimits_{n\in S} \mathbf{x}_t^n z_t^n \tag{9}$$

where S is the set that represents the top $j\%$ highest fitness score hypotheses.

In this work, the BASD hand tracker is proposed to track both hands sequentially. Algorithm 1 is first performed to track the right hand. After the location of the right hand in all frames have been obtained, the right hand region is filled with black pixels. Subsequently, Algorithm 1 is repeated to acquire the location of the left hand.

Table 1
List of isolated ASL.

Database	Isolated signs
RWTH-BOSTON-50	ariv1, bdown, box, futue, have, house, movie, pepol, toy, book, frend, hmwrk, give, leg, write

4. Experiments and discussions

To evaluate the performance of the proposed BASD hand tracker on isolated sign language recognition, experiments are carried out on the RWTH-BOSTON-50 database by Zahedi et al. [25], which is published by The National Center for Sign Language and Gesture Resources of the Boston University. This database contains 50 isolated American Sign Language and the videos were recorded at 30 frames per second with dimension of 312 × 214 pixel. In this work, only the videos acquired by the front camera are used. Table 1 enumerates the isolated sign labels considered in the experiments.

In this work, the objects that we intend to track are the right and left hands. We formulate the problem sequentially instead of tracking both hands concurrently. For comparison, we compare the proposed BASD Hand Tracker's performances with recent state-of-the-art trackers using the same initial position of the target: Incremental Visual Tracker (IVT) [26], Sparse Prototypes Tracker (SP) [27], Consistent Low-rank Sparse Tracker (LRT) [28], and Serial Particle Filter (SPF) [23]. Particle filter has been extensively used in many object tracking works [6,29,30,31,32,33]. A particle filter is described by two state equations, which are the dynamic model and observation model. IVT utilizes a novel incremental PCA approach to learn an updatable subspace representation online generatively, whereas SP introduces sparsity and trivial templates into generative PCA subspace learning to explicitly handle occlusion and motion blur. LRT formulates tracking problem as searching for the best image regions which are similar to the tracked targets by introducing sparse linear representation. SPF utilizes (1) as the dynamic model and (4) as the observation model for the particle filter.

All the experiments are executed in Matlab R2015a on the same machine with Intel(R) Xeon(R) E3-1231 processor. Before starting out on the actual experiments, the choice of the free parameters needs to be determined. The optimal value for the free parameters is determined based on the average COL errors. Keeping all other parameters constant, analyses are made by varying the value of these parameters. The parameter tuning is examined in a grid search in a promising range, as depicted in Fig. 4. Choosing a proper γ is important in characterizing the age function. γ is set

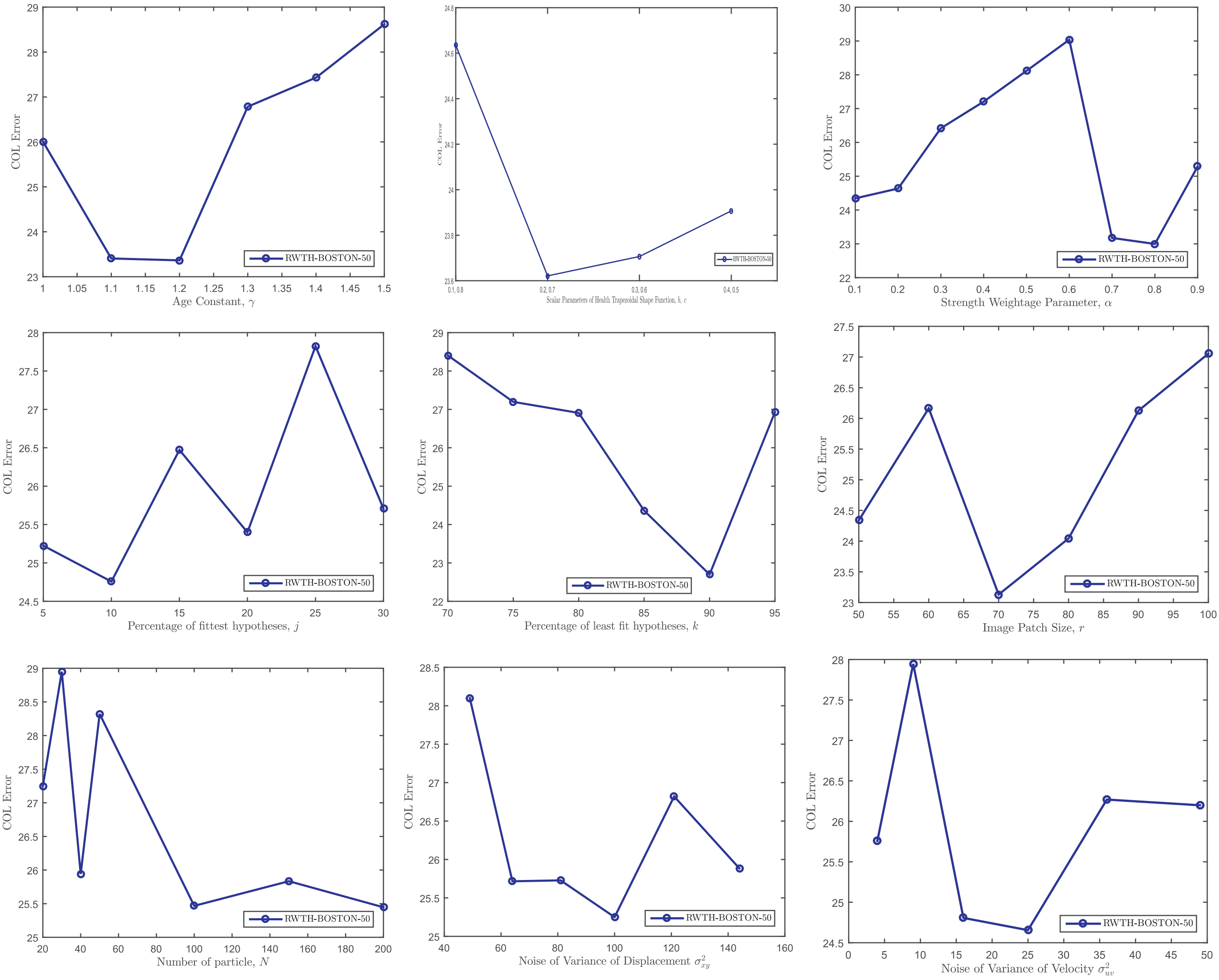


Fig. 4. COL errors for parameters setting.

Table 2
Parameter setting.

Parameter	Description	Default value
γ	Age constant	1.2
(a, b, c, d)	Scalar parameters of the trapezoidal shaped health function	(0,0.2,0.7,1)
α	Weightage parameter	0.8
j	Percentage of fittest hypotheses that is used for state estimation at each iteration	10
k	Percentage of least fit hypotheses that are removed at each iteration	90
r	Image patch size	70
N	Number of hypothesis	100
σ^2_{xy}	Noise variance of displacement	100
σ^2_{uv}	Noise variance of velocity	25

Table 3
Comparison of TER for right hand.

Isolated sign	Methods														
	BASD			IVT			SP			LRT			SPF		
	τ=10	τ=15	τ=20	τ=10	τ=15	τ=20	τ=10	τ=15	τ=20	τ=10	τ=15	τ=20	τ=10	τ=15	τ=20
ariv1	22.22	11.11	0.00	44.44	44.44	44.44	44.44	33.33	33.33	88.89	77.78	55.56	44.44	22.22	0.00
bdown	42.86	28.57	14.29	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	28.57	28.57	21.43
box	72.73	18.18	9.09	63.64	63.64	63.64	63.64	63.64	63.64	90.91	90.91	90.91	90.91	63.64	54.55
futue	87.50	87.50	81.25	0.00	0.00	0.00	6.25	0.00	0.00	93.75	93.75	87.50	56.25	31.25	12.50
have	25.00	0.00	0.00	62.50	0.00	0.00	62.50	0.00	0.00	100.00	87.50	50.00	75.00	50.00	50.00
house	80.00	70.00	50.00	65.00	60.00	55.00	70.00	60.00	55.00	95.00	95.00	95.00	75.00	70.00	60.00
movie	25.00	25.00	0.00	100.00	100.00	25.00	100.00	100.00	25.00	100.00	100.00	75.00	50.00	0.00	0.00
pepol	63.64	45.45	18.18	0.00	0.00	0.00	0.00	0.00	0.00	90.91	81.82	63.64	54.55	54.55	54.55
toy	28.57	0.00	0.00	100.00	28.57	14.29	100.00	28.57	14.29	100.00	85.71	85.71	71.43	14.29	0.00
book	87.50	87.50	25.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	87.50	37.50
frend	70.59	47.06	47.06	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	76.47	41.18	41.18
hmwrk	53.85	15.38	7.69	69.23	53.85	30.77	69.23	69.23	69.23	92.31	92.31	92.31	38.46	23.08	0.00
give	76.92	69.23	50.00	57.69	57.69	50.00	69.23	46.15	38.46	96.15	96.15	92.31	92.31	76.92	53.85
leg	80.00	36.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	92.00	80.00	68.00	36.00	28.00	28.00
write	68.00	44.00	20.00	96.00	96.00	96.00	80.00	60.00	36.00	96.00	96.00	96.00	48.00	16.00	8.00
Average	**58.96**	**39.00**	**21.50**	63.90	53.61	45.28	64.35	50.73	42.33	95.73	91.80	83.46	62.49	40.48	28.10

to a value slightly greater than 1 to demonstrate that a younger hypothesis is fitter than an older one. Similarly, α is a value to determine the strength combination of a hypothesis in the current and previous iteration. *a, b, c, d* are the four scalar parameters of the trapezoidal shape function that determines the sickness of a hypothesis. In particular, *a, d* represent the start and end of the hypothesis, whereas *b, c* represent infancy and old age in the life cycle of the hypothesis. The optimal setting of the parameters in the proposed method is listed in Table 2.

To evaluate the hand tracking methods in a comprehensive way, two evaluation methods are considered, namely quantitative evaluation and qualitative evaluation. The same evaluation methods are used by many object tracking researches [34,35]. The trackers are evaluated quantitatively both in terms of Tracking Error Rate (TER) proposed by Dreuw et al. [36], and a generic Center-of-location (COL) error. Let $\mathbf{g}_t$ denotes the annotated groundtruth hand positions of tth frame, and τ is the threshold. The TER is given by:

$$TER = \frac{1}{T}\sum_{t=1}^{T}\delta_\tau(\mathbf{g}_t, \mathbf{x}_t) \quad \text{with}$$

$$\delta_\tau(\mathbf{g}, \mathbf{x}) = \begin{cases} 0 & \text{if } \parallel \mathbf{g} - \mathbf{x} \parallel < \tau \\ 1 & \text{otherwise} \end{cases} \tag{10}$$

COL error calculates the Euclidean distance in pixels between the centre of the tracked hand patches and the centre of the groundtruth hand patches. For qualitative evaluation, some illustrated tracking results are shown and analyzed. The trajectories obtained by the proposed BASD, IVT, SP, LRT and SPF are compared to the groundtruth trajectory which was acquired by visual inspection.

TER is employed to measure the distance for both hands of all trackers to the groundtruth for $\tau = 10, 15, 20$. The results of both hands are shown in Table 3 and Table 4, respectively. Additionally, Tables 5 and 6 show the results in terms of COL error for both hands. From the experimental results, it is notable that BASD performs better than IVT, SP, LRT and SPF. Among the set of methods tested, BASD method achieves the lowest COL errors.

The hand tracking execution time which measured in number of frame per second of all methods is presented in Table 7. The particle filter based methods, i.e., IVT, SP, and LRT consume higher execution time by about 6 fps to 7 fps. The execution time of proposed BASD method is slightly lower than particle filter based methods. The computation time of particle filter based methods and BASD is highly dependent on the number of particle/hypothesis and the complexity of the observational model. IVT, SP and LRT consume more computational time due to their complex update model for the tracked object. On average, BASD achieves either the best or second best performance in most isolated signs. Therefore, we conclude that the trajectories obtained by BASD are relatively closer to the groundtruth trajectories, and hence, demonstrates that BASD can better detect the trajectory of the hands.

Additionally, we also evaluate the hand tracking qualitatively. We choose a number of isolated signs for visual inspection. Fig. 5 shows the tracking result of isolated signs *ariv1, bdown, hmwrk, toy* and *write*. Noticeably, the proposed BASD method performs better than IVT, SP, LRT and SPF in all test cases. The performance of SPF highly depends on its two state equations, and all candidates in SPF are resampled at every iteration. In contrast, BASD mimics the life cycle of human and updates and enhances the hypothesis at every iteration based on its fitness which is, in turn, affected

Table 4
Comparison of TER for left hand.

Isolated sign	Methods														
	BASD			IVT			SP			LRT			SPF		
	t=10	*t*=15	*t*=20	*t*=10	*t*=15	*t*=20	*t*=10	*t*=15	*t*=20	*t*=10	*t*=15	*t*=20	*t*=10	*t*=15	*t*=20
ariv1	55.56	44.44	11.11	66.67	55.56	44.44	66.67	55.56	44.44	100.00	88.89	88.89	66.67	44.44	33.33
bdown	78.57	42.86	28.57	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	64.29	50.00
box	72.73	54.55	27.27	72.73	72.73	72.73	72.73	72.73	72.73	90.91	90.91	90.91	81.82	63.64	27.27
futue	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	81.25	93.75	93.75	93.75	75.00	56.25	31.25
have	25.00	12.50	12.50	0.00	0.00	0.00	0.00	0.00	0.00	87.50	87.50	87.50	37.50	25.00	12.50
house	90.00	85.00	85.00	80.00	70.00	60.00	80.00	70.00	60.00	95.00	95.00	95.00	95.00	65.00	50.00
movie	75.00	50.00	0.00	50.00	0.00	0.00	50.00	0.00	0.00	75.00	75.00	75.00	0.00	0.00	0.00
pepol	36.36	18.18	9.09	27.27	9.09	0.00	18.18	9.09	0.00	90.91	90.91	90.91	72.73	54.55	54.55
toy	14.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	85.71	85.71	85.71	42.86	14.29	0.00
book	100.00	87.50	50.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	62.50
frend	94.12	88.24	82.35	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
hmwrk	76.92	76.92	53.85	15.38	7.69	7.69	7.69	0.00	0.00	100.00	92.31	92.31	92.31	46.15	38.46
give	65.38	50.00	23.08	88.46	84.62	80.77	88.46	84.62	80.77	96.15	96.15	96.15	73.08	73.08	69.23
leg	100.00	80.00	60.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
write	92.00	80.00	60.00	96.00	92.00	88.00	92.00	92.00	92.00	96.00	96.00	96.00	84.00	72.00	60.00
Average	**70.48**	**56.76**	**38.94**	65.18	58.20	55.66	63.80	57.68	55.41	94.06	92.81	92.81	74.73	58.58	45.94

Table 5
COL errors in pixels of RWTH-Boston-50 right hand.

RWTH-BOSTON-50	BASD	IVT	SP	LRT	SPF
ariv1	5.00	4.87	4.92	35.66	5.86
bdown	9.55	28.84	29.09	86.63	9.28
box	5.48	50.96	51.27	114.70	6.21
futue	2.39	4.01	3.84	68.12	8.78
have	9.21	9.31	9.31	48.34	17.71
house	3.70	22.10	22.51	98.84	14.39
movie	11.41	14.06	14.06	96.92	3.80
pepol	4.20	3.13	2.80	74.43	4.46
toy	81.54	12.66	13.07	61.81	81.90
book	11.92	34.75	34.75	109.43	11.03
frend	7.07	48.13	48.23	22.11	23.03
hmwrk	9.02	14.95	13.78	105.01	9.63
give	7.30	26.33	29.29	52.60	17.04
leg	2.58	4.67	5.24	75.93	4.74
write	4.07	41.14	41.48	119.94	7.55
Average	**11.63**	21.33	21.58	78.03	15.03

Table 6
COL errors in pixels of RWTH-Boston-50 left hand.

RWTH-BOSTON-50	BASD	IVT	SP	LRT	SPF
ariv1	10.73	17.99	15.07	52.93	11.88
bdown	9.99	50.03	49.60	152.92	17.40
box	11.33	47.89	47.79	222.08	9.94
futue	4.60	45.52	45.34	91.23	5.26
have	2.62	5.18	4.92	126.30	6.38
house	9.05	25.54	27.92	137.93	33.62
movie	35.18	6.51	6.32	92.23	13.79
pepol	6.60	7.32	7.14	166.84	5.78
toy	3.87	5.55	5.57	164.53	4.13
book	25.93	31.58	30.84	128.35	28.74
frend	50.85	41.87	41.69	85.12	33.90
hmwrk	6.82	34.18	33.71	94.07	12.57
give	40.12	56.18	43.02	94.73	31.83
leg	6.21	68.28	68.78	66.10	23.88
write	20.90	24.86	23.27	123.29	21.27
Average	**16.32**	31.23	30.07	119.91	17.36

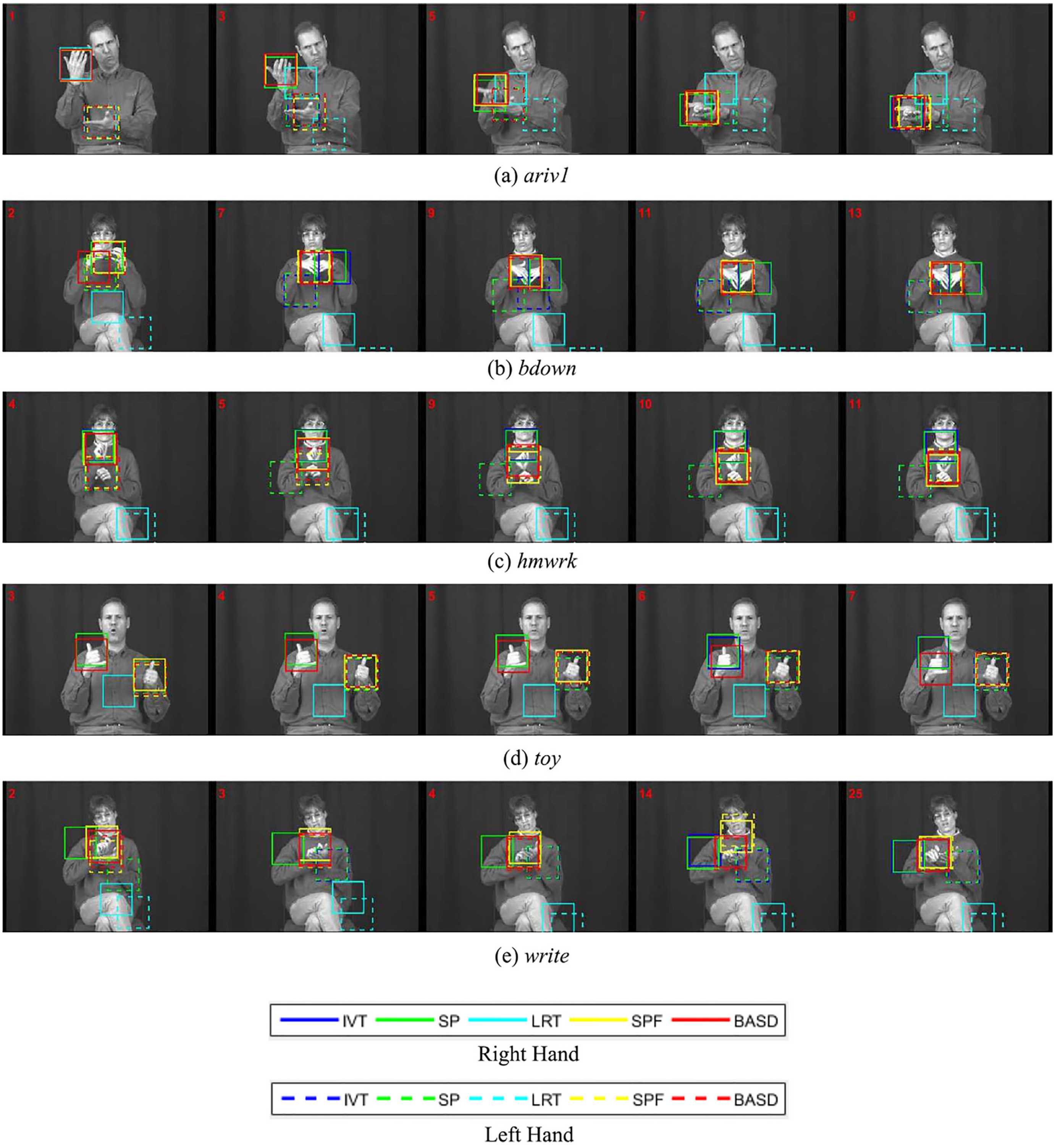
Table 7
The execution time of hand tracking in number of frame per second (fps).

Method	BASD	IVT	SP	LRT	SPF
Execution time (fps)	7.64	6.50	6.30	5.70	31.50

by the strength and age of the hypothesis. As for IVT, SP and LRT, they perform poorly because their performance highly depends on holistic representations of the first frame and they are not robust against partial occlusion.

5. Conclusion

In this paper, a hand tracking framework inspired by the four dukkha (birth, aging, sickness and death) in Buddha's teaching is



(a) *ariv1*

(b) *bdown*

(c) *hmwrk*

(d) *toy*

(e) *write*

Fig. 5. Tracking results of (a) *ariv1*, (b) *bdown*, (c) *hmwrk*, (d) *toy* and (e) *write*.

proposed. The proposed BASD hand tracking framework resembles a standard state-space model, but maintains multiple hypotheses. Hypothesis update and propagation mechanisms that adopt the idea of BASD is incorporated into the proposed hand tracking framework. Hand motion in sign language video is first modelled using a hypothesis update function. The hypothesis is then evaluated by a fitness function which is computed based on the aging, sickness and strength of the hypothesis. Hypotheses with low fitness score are eliminated, and new hypotheses are birthed at the midpoint of the surviving parent hypotheses. From the experimental results, the proposed BASD hand tracker is able to track the hands accurately and promising results are obtained.

References

[1] F.-S. Chen, C.-M. Fu, C.-L. Huang, Hand gesture recognition using a real-time tracking method and hidden Markov models, Image Vis. Comput. 21 (8) (2003) 745–758.

[2] S.S. Rautaray, A. Agrawal, A real time hand tracking system for interactive applications, Int. J. Comput. Appl. 18 (6) (2011) 28–33.

[3] Z. Zhang, F. Huang, Hand tracking algorithm based on superpixels feature, in: Proceedings of the 2013 International Conference on Information Science and Cloud Computing Companion (ISCC-C), IEEE, 2013, pp. 629–634.

[4] Y.F.A. Gaus, F. Wong, Hidden Markov model-based gesture recognition with overlapping hand-head/hand-hand estimated using Kalman Filter, in: Proceedings of the 2012 Third International Conference on Intelligent Systems, Modelling and Simulation (ISMS), IEEE, 2012, pp. 262–267.

[5] S. Park, S. Yu, J. Kim, S. Kim, S. Lee, 3D hand tracking using Kalman filter in depth space, EURASIP J. Adv. Sig. Process. 2012 (1) (2012) 1–18.

[6] C. Shan, T. Tan, Y. Wei, Real-time hand tracking using a mean shift embedded particle filter, Pattern Recognit. 40 (7) (2007) 1958–1970.

[7] S. Belgacem, C. Chatelain, A. Ben-Hamadou, T. Paquet, Hand tracking using optical-flow embedded particle filter in sign language scenes, in: Proceedings of the International Conference on Computer Vision and Graphics, Springer, 2012, pp. 288–295.

[8] P. Campr, M. Hrúz, A. Karpov, P. Santemiz, M. Železnỳ, O. Aran, Sign-language-enabled information kiosk, eNTERFACE'08 (2009) 24–33.

[9] M. Morshidi, T. Tjahjadi, Gravity optimised particle filter for hand tracking, Pattern Recognit. 47 (1) (2014) 194–207.

[10] Y. Jang, et al., Gesture recognition using depth-based hand tracking for contactless controller application, in: Proceedings of the 2012 IEEE International Conference on Consumer Electronics (ICCE), 2012, pp. 297–298.

[11] B. Yoo, J.-J. Han, C. Choi, K. Yi, S. Suh, D. Park, C. Kim, 3D user interface combining gaze and hand gestures for large-scale display, in: CHI'10 Extended Abstracts on Human Factors in Computing Systems, ACM, 2010, pp. 3709–3714.

[12] M. Kolsch, M. Turk, Fast 2d hand tracking with flocks of features and multi-cue integration, in: Proceedings of the 2004 International Conference on Computer Vision and Pattern Recognition Workshop, IEEE, 2004, p. 158.

[13] C. Chen, M. Zhang, K. Qiu, Z. Pan, Real-time robust hand tracking based on CAMSHIFT and motion velocity, in: Digital Home (ICDH), 2014 5th International Conference on, IEEE, 2014, pp. 20–24.

[14] C.-P. Chen, Y.-T. Chen, P.-H. Lee, Y.-P. Tsai, S. Lei, Real-time hand tracking on depth images, in: Visual Communications and Image Processing (VCIP), IEEE, 2011, pp. 1–4.

[15] M. Park, M.M. Hasan, J. Kim, O. Chae, Hand detection and tracking using depth and color information, in: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition, vol. 2, 2012, pp. 779–785.

[16] P. Kishore, D.A. Kumar, E. Goutham, M. Manikanta, Continuous sign language recognition from tracking and shape features using fuzzy inference engine, in: Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking, IEEE, 2016, pp. 2165–2170.

[17] P. Jangyodsuk, C. Conly, V. Athitsos, Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features, in: Proceedings of the Seventh International Conference on PErvasive Technologies Related to Assistive Environments, ACM, 2014, p. 50.

[18] J. Zhang, W. Zhou, H. Li, A new system for chinese sign language recognition, in: Proceedings of the 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), IEEE, 2015, pp. 534–538.

[19] C. Sun, T. Zhang, C. Xu, Latent support vector machine modeling for sign language recognition with kinect, ACM Trans. Intell. Syst. Technol.(TIST) 6 (2) (2015) 20.

[20] T. Liu, W. Zhou, H. Li, Sign language recognition with long short-term memory, in: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 2871–2875.

[21] B. Bodhi, The Noble Eightfold Path–The Way to the End of Suffering, The Wheel Publication, 1984.

[22] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, IEEE, 2001, pp. I–511.

[23] K.M. Lim, A.W. Tan, S.C. Tan, A feature covariance matrix with serial particle filter for isolated sign language recognition, Expert Syst. Appl. 54 (2016) 208–218.

[24] N. Otsu, A threshold selection method from gray-level histograms, Automatica 11 (285–296) (1975) 23–27.

[25] M. Zahedi, D. Keysers, T. Deselaers, H. Ney, Combination of tangent distance and an image distortion model for appearance-based sign language recognition, in: Pattern Recognition, Springer, 2005, pp. 401–408.

[26] D.A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (1–3) (2008) 125–141.

[27] D. Wang, H. Lu, M.-H. Yang, Online object tracking with sparse prototypes, IEEE Trans. Image Process. 22 (1) (2013) 314–325.

[28] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, B. Ghanem, Robust visual tracking via consistent low-rank sparse learning, Int. J. Comput. Vis. 111 (2) (2015) 171–190.

[29] K. Nummiaro, E. Koller-Meier, L. Van Gool, An adaptive color-based particle filter, Image Vis. Comput. 21 (1) (2003) 99–110.

[30] K. Okuma, A. Taleghani, N. De Freitas, J.J. Little, D.G. Lowe, A boosted particle filter: multitarget detection and tracking, in: Computer Vision-ECCV 2004, Springer, 2004, pp. 28–39.

[31] C. Chang, R. Ansari, Kernel particle filter for visual tracking, IEEE Signal Process. Lett. 12 (3) (2005) 242–245.

[32] C. Yang, R. Duraiswami, L. Davis, Fast multiple object tracking via a hierarchical particle filter, in: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV), vol. 1, IEEE, 2005, pp. 212–219.

[33] K. Nummiaro, E. Koller-Meier, L. Van Gool, Object tracking with an adaptive color-based particle filter, in: Pattern Recognition, Springer, 2002, pp. 353–360.

[34] A.W. Smeulders, D.M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, M. Shah, Visual tracking: an experimental survey, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1442–1468.

[35] Y. Wu, J. Lim, M.-H. Yang, Object tracking benchmark, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1834–1848.

[36] P. Dreuw, J. Forster, H. Ney, Tracking benchmark databases for video-based sign language recognition, in: Trends and Topics in Computer Vision, Springer, 2012, pp. 286–297.

Kian Ming Lim received his bachelor's degree in Information Systems Engineering and M.Eng.Sc. from Multimedia University, Malaysia, in 2004 and 2011, respectively. He is currently working toward PhD degree in the area of computer vision at Faculty of Information Science and Technology, Multimedia University, Malaysia. His research interests include computer vision, pattern classification and machine learning.

Dr. Alan Tan received the B.Eng. (Hons) degree in Electrical Engineering from University of Malaya, Kuala Lumpur, in 1999 on the Shell scholarship, and the M.EngSc. and Ph.D. degrees from Multimedia University, Selangor, in 2003 and 2008, respectively. He is currently an Associate Professor, and the Dean of the engineering faculty at Multimedia University. His research interests include pattern recognition, signal processing, and wireless communications systems.

Dr. Shing Chiang Tan received his B.Tech. (Hons.) and M.Sc. (Eng.) degrees from Universiti Sains Malaysia. He obtained his Ph.D. degree from Multimedia University, Malaysia. He is currently an Associate Professor with the Faculty of Information Science and Technology, Multimedia University. His current research interests include computational intelligence and its applications, which include but are not limited to, data classification, pattern recognition, industrial process condition monitoring, fault detection and diagnosis, optimisation and biomedical disease data analysis and classification. Dr. Tan was a recipient of the Matsumae International Foundation Fellowship, Japan, in 2010.