Machine Vision and Applications (2018) 29:233–245
https://doi.org/10.1007/s00138-017-0896-5

ORIGINAL PAPER

Exercise classification and event segmentation in Hammersmith Infant Neurological Examination videos

Abdul Fatir Ansari[1] · **Partha Pratim Roy**[2] · **Debi Prosad Dogra**[3]

Received: 19 June 2016 / Revised: 21 May 2017 / Accepted: 29 November 2017 / Published online: 8 December 2017

Abstract
Image and video processing techniques are being frequently used in medical science applications. Computer vision-based systems have successfully replaced various manual medical processes such as analyzing physical and biomechanical parameters, physical examination of patients. These systems are gaining popularity because of their robustness and the objectivity they bring to various medical procedures. Hammersmith Infant Neurological Examinations (HINE) is a set of physical tests that are carried out on infants in the age group of 3–24 months with neurological disorders. However, these tests are graded through visual observations, which can be highly subjective. Therefore, computer vision-aided approach can be used to assist the experts in the grading process. In this paper, we present a method of automatic exercise classification through visual analysis of the HINE videos recorded at hospitals. We have used scale-invariant-feature-transform features to generate a bag-of-words from the image frames of the video sequences. Frequency of these visual words is then used to classify the video sequences using HMM. We also present a method of event segmentation in long videos containing more than two exercises. Event segmentation coupled with a classifier can help in automatic indexing of long and continuous video sequences of the HINE set. Our proposed framework is a step forward in the process of automation of HINE tests through computer vision-based methods. We conducted tests on a dataset comprising of 70 HINE video sequences. It has been found that the proposed method can successfully classify exercises with accuracy as high as 84%. The proposed work has direct applications in automatic or semiautomatic analysis of "vertical suspension" and "ventral suspension" tests of HINE. Though some of the critical tests such as "pulled-to-sit," "lateral tilting," or "adductor's angle measurement" have already been addressed using image- and video-guided techniques, scopes are there for further improvement.

Keywords HINE tests · Exercise classification · Video segmentation · Bag-of-words · Event segmentation

1 Introduction

Medical research has seen huge advancement in the field of computer vision-guided automatic and semiautomatic systems. Image and video outputs from X-ray, electrocardiography (ECG), electroencephalography (EEG), ultrasound (USG), and magnetic resonance (MR) are analyzed by physicians with the help of computers to make the diagnostic process swift and efficient. These are used mainly to study and understand the internal structures of human body. External imaging using vision sensors such as camera can act as an important diagnostic or maintenance utility. External imaging has been used in human gait analysis [1], patient surveillance [2], pedestrian detection [3], infant or old person monitoring systems [4,5], etc.

Researchers have shown that analysis of Hammersmith Infant Neurological Examinations (HINE) can be automated using computer vision-guided tools. HINE is a diagnostic process that is often used for the assessment of neurological development of infants under the age group of 2–24 months [6]. Neurological diseases in infants are common, and often

✉ Debi Prosad Dogra
dpdogra@iitbbs.ac.in

Abdul Fatir Ansari
abdulfatir@outlook.com

Partha Pratim Roy
proy.fcs@iitr.ac.in

[1] Department of Civil Engineering, IIT Roorkee, Roorkee 247667, India

[2] Department of Computer Science and Engineering, IIT Roorkee, Roorkee 247667, India

[3] School of Electrical Sciences, IIT Bhubaneswar, Bhubaneswar 751013, India

such diseases hamper a child's mental development. HINE tests are efficient in assessment of the possibility of neurological diseases in infants that can be prevented using suitable precautionary measures. Survival rate of preterm and low birth-weight newborns can be increased if diagnosis of such disorders is done at early stage [7]. HINE tests include assessment of posture, movements, cranial nerve functions, tone, reflexes, and behavior. These tests are carried out by visually observing the reactions of the infant during various exercises. The tests, however, as it has been learned from experts of this domain, are subjective. Therefore, it is believed and has already been proven to some extent by various researchers that computer vision- aided methods can substantially help in providing the necessary objectivity to the tests [8–11]. However, some of the critical tests have yet not been tried for objective evaluation using computer vision-guided techniques. Therefore, we feel, a generic model to segment events (individual tests) can be helpful for developing fully automatic HINE grading.

In this paper, we present an efficient method of exercise classification and event segmentation in video sequences of the HINE exercises. We use scale-invariant-feature-transform (SIFT) features to generate a bag-of-words for frames in the video sequences. We then use the frequency of these visual words to classify the video sequences using an HMM-based two-pass classifier. We also present a method of event segmentation in long videos that contain two to three exercises of the HINE set. It has been found that the proposed method can successfully classify exercises with accuracy as high as 84%. Such event segmentation and exercise classification is a step forward in automation of HINE tests. Using the methods proposed in this paper, one can segment and classify long videos of HINE exercise sets into videos of individual exercises. The smaller videos can then be used for automatic analysis of individual exercises. Video recordings of the examinations are performed in controlled environments. Infants of various ages, sizes, and complexions undergo these tests. Camera placement and orientation also affect outcome. These factors make the problem of exercise classification a challenging task, and a robust classifier is therefore required. Visual similarity of different exercises also makes the event segmentation task more challenging than it appears to be.

The rest of the paper is organized as follows: Information about the work related to our proposed work is presented in Sect. 2. Our proposed framework is explained in Section 3. Details about the experiments and results are demonstrated in Sect. 4. A comparative analysis with existing classifiers is presented in Sect. 4.4. Limitations of the proposed framework are explained in Sect. 5. We present conclusion and possible extensions of our work in Sect. 6.

2 Related work

Romeo et al. [7] have experimentally validated that integration of traditional neurological tests with general movements can be used for neurological assessment in preterm infants. They have also shown that neuromotor development in infants with cerebral palsy (CP) can be investigated using HINE during the first 12 months of the age of the infant [12]. However, as mentioned earlier, grading of these tests is highly subjective. Therefore, researchers have already proposed computer vision techniques for assessment of a few of the tests of HINE.

Dogra et al. have proposed computer vision-guided techniques for the assessment of some of the critical tests of the HINE set, namely lateral tilting [9], adductor's angle measurement [8,10], and pulled-to-sit [11] as shown in Fig. 1. Grading of vertical suspension proposed by Dey et al. [13] is another example of the use of computer vision in healthcare domain.

The problem of efficient classification of exercises and event segmentation, however, still exists. Event segmentation and exercise classification can prove to be a step forward in complete automation of many of the critical tests. It has been observed that videos of these tests are usually recorded in a continuous manner. Therefore, segmentation of the continuously recorded video and classification of the tests remain two important problems to be addressed. Since some of these tests look similar in visual context, an accurate classification will essentially help us to build systems for objective evaluation. This has been proposed in this work.

Such automatic computer-based systems can be quite helpful in NICU environment. For example, the web-enabled health information system for neonatal intensive care unit proposed by Roy et al. [14] is presently being used in NICU of SSKM Hospital, Kolkata. Dogra et al. [15] have developed a tool for recording HINE scores inside neurodevelopment clinic of the hospitals.

In [16], we have already proposed a preliminary approach of classification with the help of skeleton bounding box features using hidden Markov model-based sequential classifier applied on a small dataset of 25 videos. We have achieved an average accuracy of 78%. This has motivated us to look deeper into the concept and propose robust features that can be successfully used for more accurate classification on a larger dataset.

3 Proposed framework

The schematic diagram of the proposed framework is shown in Fig. 2. Our objective is to classify different exercises in the video sequences of the HINE set. The four classes in our dataset are named: (a) *pulled-to-sit*, (b) *vertical suspension*,

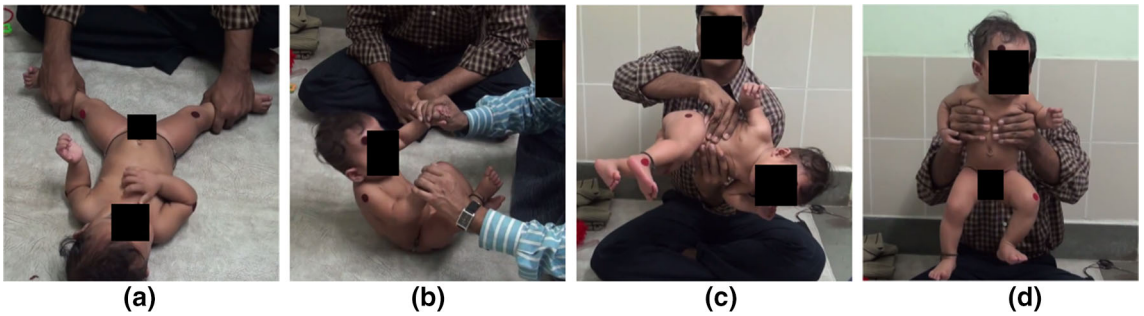

Fig. 1 Four exercises of HINE set. **a** Adductor's angle, **b** pulled-to-sit, **c** lateral tilting, and **d** vertical suspension

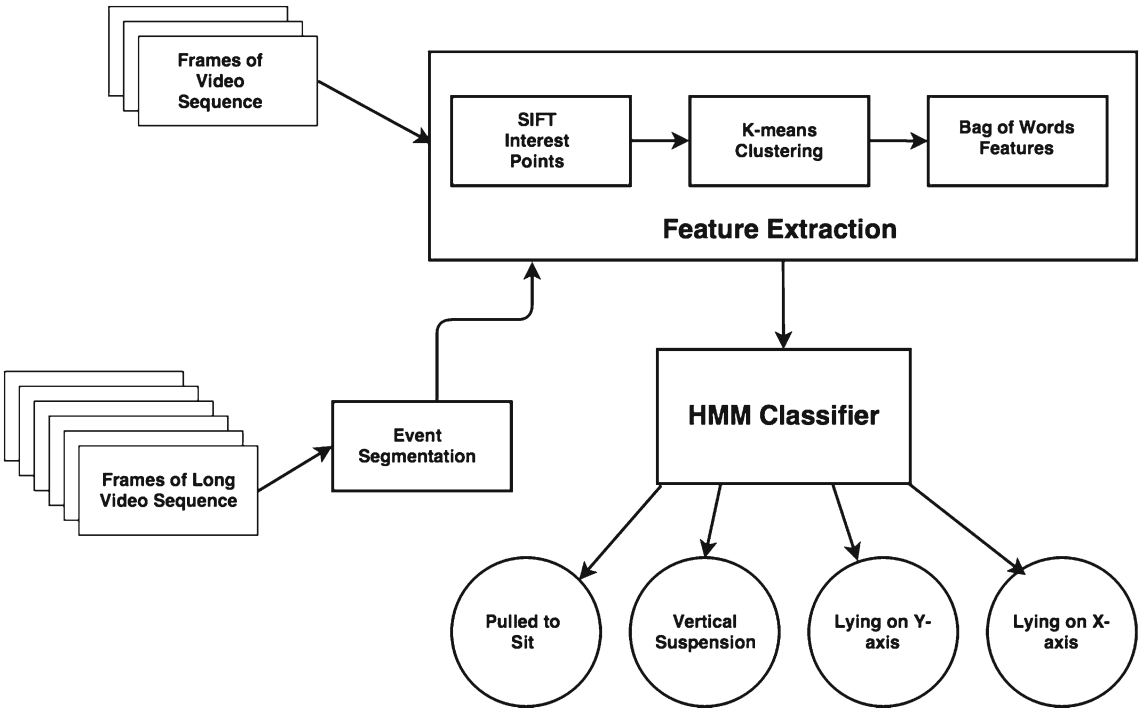

Fig. 2 Functional block diagram of the proposed methodology

(c) *laying-horizontally*, and (d) *laying-vertically*. In *pulled-to-sit*, the infant is made to sit by pulling through its hands. In *Vertical Suspension*, the infant is raised and suspended vertically. In *laying-horizontally* and *laying-vertically*, the infant lies along the horizontal and vertical axes, respectively. The four classes are shown in Fig. 3. Each sequence is represented by a number of descriptors equal to the number of frames in the sequence. Choice of features is important for sequence classification because it is desirable to know how the features evolve over time for a given sequence. After the extraction of features from the video sequences, we have performed HMM-based classification using a two-pass algorithm.

We also present a method of event segmentation in long HINE video sequences containing more than one classes or exercises. Initially, the magnitude of optical flow across all frames in the video sequence is calculated. Next, abrupt and large changes in optical flow magnitude are detected. These points of sudden change are marked as change-event (exercise) points. This helps in automatic indexing [17] of exercises in video sequences.

3.1 Feature extraction

The bag-of-words approach is a widely used technique applied for object recognition [18]. The idea behind the approach is to represent an image by a histogram of visual words where each word corresponds to a local interest point extracted from the image. In the proposed work, we process a video frame by frame to extract the scale-invariant-feature-transform (SIFT) [19] descriptor. SIFT interest points in a sample frame from each of the aforementioned four classes are shown in Fig. 3. The detected interest points from a large

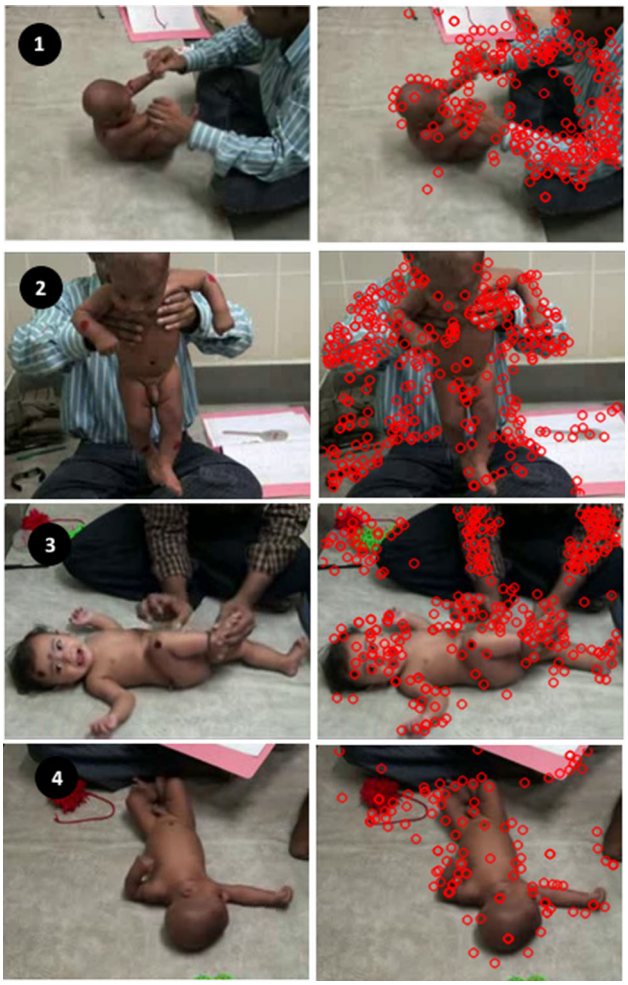

Fig. 3 SIFT interest points on the four different exercise classes (1) pulled-to-sit, (2) vertical suspension, (3) laying-horizontally, and (4) laying-vertically

number of frames are clustered using k-means algorithm to generate a codebook. Then, we extract a descriptor with size equal to the codebook. This contains the frequency with which each visual word occurs in the video frame. Such a representation allows us to analyze how specific words appear and disappear during the course of a sequence giving the pattern a unique characteristic.

Since some of the visual words occur naturally in large frequencies in most of the frames, we perform the term frequency-inverse document frequency (tf-idf) [20,21] vectorization of the descriptor to account for large occurrences of these visual words. For that we calculate the tf-idf weight ω as shown in Eq. 1.

$$\omega = f_{t,d} * log \frac{N}{n_t} \tag{1}$$

where $f_{t,d}$ represents the number of times a term t occurs in a document d, N represents the total number of documents, and n_t is the number of documents where the term t appears. This step generates a feature vector of size k for each frame and n such feature vectors for each video sequence, where n is the total number of frames in the video.

Two-pass algorithm During the course of experiments, it has been observed that the single-pass algorithm—in which only bag-of-words features are used for classification—created confusion between two classes, namely *pulled-to-sit* and *laying-vertically*. Hence, we have decided to merge these two classes and perform the first pass of HMM classification. Then, we have extracted features that helped us to distinguish between these two classes in the second pass of classification (Fig. 4).

During the second pass, each frame of the video sequence is smoothed using a 5×5 mean filter and then converted to YC_rC_b color space. We have tested various color spaces in our previous work [16] and found the YC_rC_b color space to be the best for skin color-based segmentation. This is because the Y or luminance component of this color space does not affect the segmentation, and hence, skins of different complexions can be easily detected. The image is then segmented based on the bounds given by Chai and Ngan [22] for detection of skin color $C_r = [140, 173]$ and $C_b = [77, 127]$. Morphological operations are performed on the output binary image to remove spurious detection. For this, morphological erosion using a 3×3 rectangular structuring element is performed on the binary image, and then, a morphological closing operation is performed using a 5×5 rectangular structuring element. The largest blob (area-wise) is chosen as the blob representing the infant's body. The rectangle bounding this blob is found out. The width, height, and aspect ratio (width/height) of the bounding box are taken as the features for the second pass of classification. We call these features *Blob Bounding Box Features*. Outputs at various stages are shown in Fig. 5.

3.2 Classification

Hidden Markov models (HMMs) have applications in temporal pattern recognition such as handwriting recognition [23], speech recognition [24], and gesture recognition [25]. Since our objective is also to perform sequence classification, HMM is considered to be a good choice given its capacity to learn from previous states. HMM can be defined by initial state probabilities P_i, state transition matrix $A = [a_{ij}]$, $i, j = 1, 2, \ldots, N,$, where a_{ij} denotes the transitions probability from state i to state j, and the output probability $b_j\ (O_k)$ is modeled with continuous output probability density functions. The density function is written as $b_j(x)$, where x represents the k-dimensional feature vector. A separate Gaussian mixture model is defined for each state. The recognition is performed using Viterbi algorithm. HTK tool-kit has been used for the implementation of the HMM classifier.

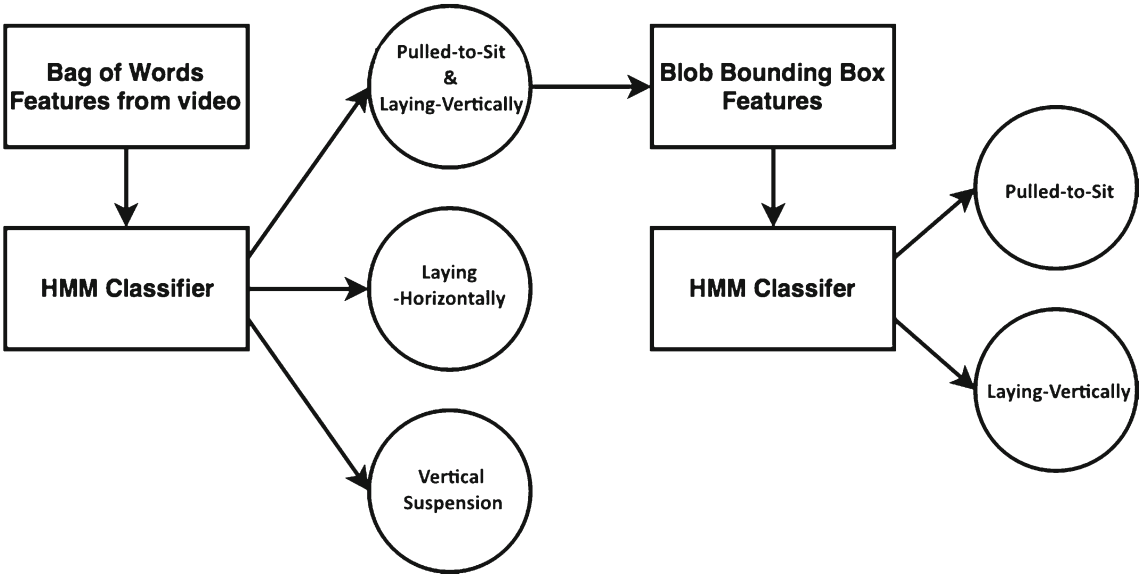

Fig. 4 Block diagram of the proposed two-pass algorithm of classification of various tests

3.3 Event segmentation in long videos

In long video sequences, which usually consist of two or more different exercises, we have used an optical flow-based algorithm to detect the temporal locations where there is a change in the exercise class. The optical flow velocities between successive frames of the video sequences have been calculated using the Lucas–Kanade algorithm [26]. The following constraint given in (2) has been solved to compute the optical flow,

$$I_{xu} + I_{yv} + I_t = 0 \tag{2}$$

where I_x, I_y, and I_t represent the spatiotemporal image brightness derivatives, u represents the optical flow in horizontal direction, and v is the optical flow in vertical direction. Lucas–Kanade method divides the original image into smaller sections to solve for u and v and assumes constant velocity in each section. Then, a weighted least-square approximation is performed to a constant model for $[u \;\; v]^T$ in each section Ω. The formula given in (3) is minimized to accomplish the above task.

$$\sum_{x \in \Omega} W^2 [I_{xu} + I_{yv} + I_t = 0]^2 \tag{3}$$

The noise threshold factor has been set to 0.009. The magnitudes of optical flow velocities for each frame are calculated and stored in an array of magnitudes. The local maximums of magnitude values are searched, and the highest ones where the difference between the peak and its neighbors is greater than 4 times the average of magnitudes of all frames in the video sequence, are chosen. The local maximums are calculated over a region of 10 frames, i.e., 5 before and 5 after the point of change in exercise. The value of *4 times the average of magnitudes* was suitably chosen after performing various experiments. A value less than 4 would lead to event segmentation even at points where there is no change in the exercise being performed. Similarly, a higher value would miss certain points where there is a change in the exercise being performed. These spikes in the magnitudes of optical flow denote the points of change event. The method is described in Algorithm 1.

Algorithm 1 Event Segmentation Algorithm

```
procedure SEGMENTEVENTS
    video ← input video sequence
    i ← frame number (0)
    magnitudes ← array of optical flow velocity magnitudes
    while hasFrame(video) do
        frame ← current frame
        framegray ← frame converted to grayscale
        flow ← optical flow of current frame using Lucas-Kanade algorithm
        magnitude ← sum of velocity magnitudes in flow
        append magnitude to magnitudes
        i = i + 1
    avg ← sum(magnitudes)/i
    findpeaks(magnitudes)                    ▷ find the highest local maximums where the difference between the peak and its neighbors is greater than 4 times of avg
```

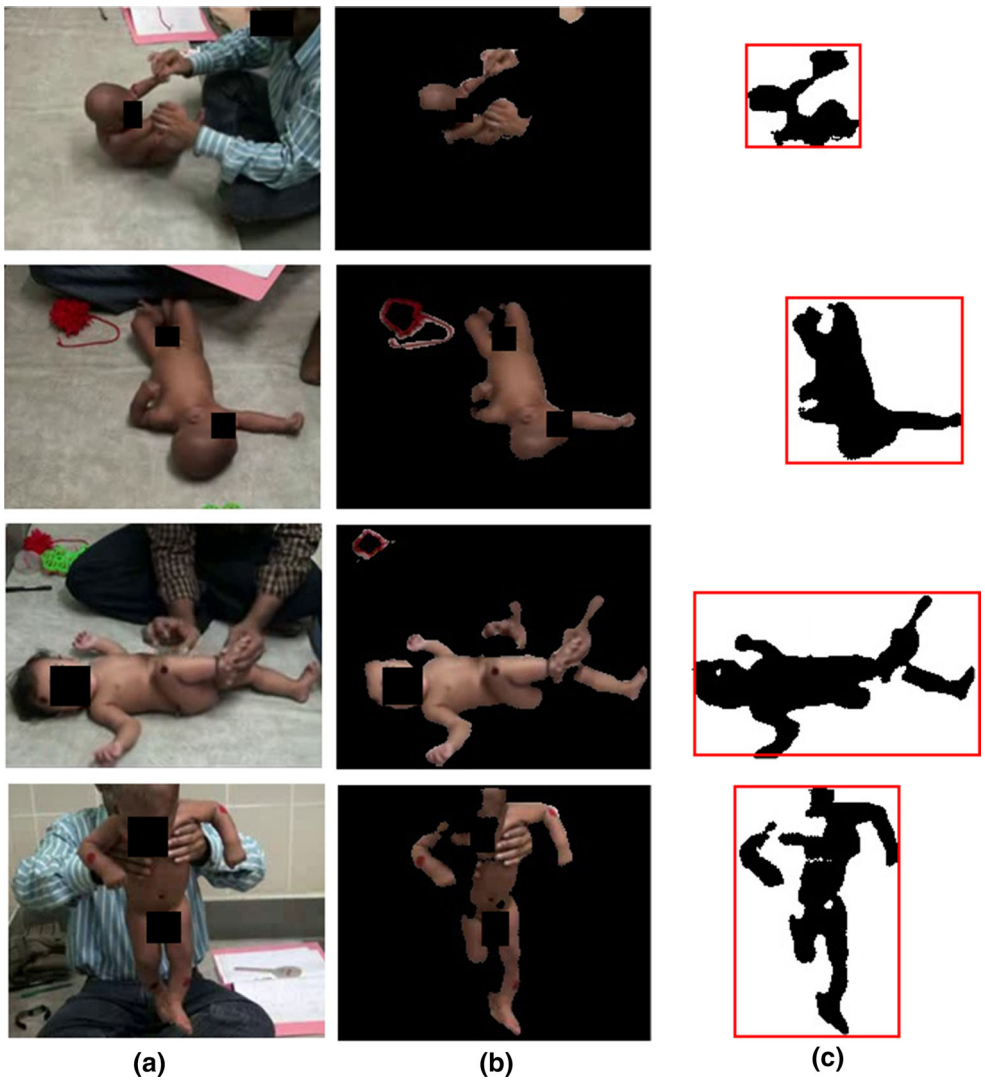

Fig. 5 Detection of Blob Bounding Box Features. **a** The input frames, **b** results after skin color-based segmentation, and **c** results after binarizing the skin color regions detected in previous step, performing morphological operations on the binary image, and choosing the largest blob area-wise

4 Results and experiments

4.1 Data details

The videos of HINE tests were recorded by Dogra et al. [8] in a controlled environment and well-established setup of neurodevelopment clinic of the Department of Neonatology of SSKM Hospital Kolkata, India. A fixed level of illumination was maintained throughout the tests. A homogeneous and contrasting background was maintained. All clothes of the subjects were removed, and the tests were conducted by experts. A single video camera was used to record the visuals of the experiments. The complete experimental setup can be found in [8]. The video clips were in AVI format at a QVGA resolution (320 × 240 pixels @ 25 fps). The dataset consists a total of 70 videos belonging to four classes, namely *Pullet-to-Sit* , *laying-horizontally*, *laying-vertically*, and *vertical suspension* with more than 10 videos of every class. The dataset consists of videos of subjects with different skin colors and age groups ranging from 3 to 24 months. We have used OpenCV toolbox to process these videos.

4.2 Results of classification

Classification using HMM has been performed on 5 videos after training the model using the remaining 65 videos. Testing has been carried out with 10 such iterations in Monte Carlo cross-validation scheme. HMM parameters have been set based on the validation set. The classifier provided best results with 4 states and 4 Gaussian mixture models (GMMs). The variation of accuracy with number of states and number of GMMs is shown in Figs. 6 and 7, respectively.

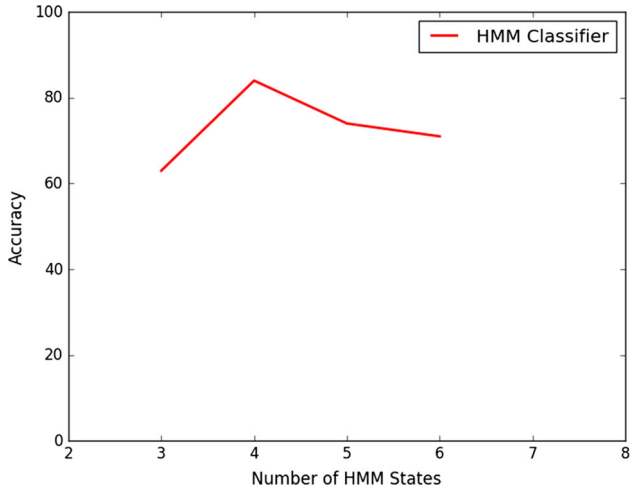

Fig. 6 Variation of accuracy with respect to number of states in the hidden Markov model (best results were obtained with 4 states)

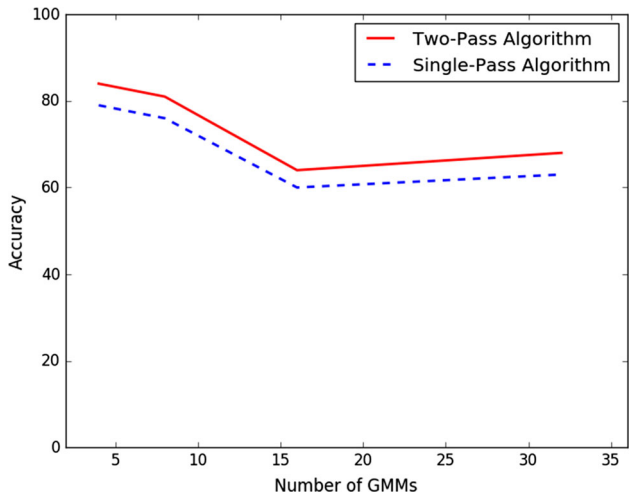

Fig. 7 Variation of accuracy with number of Gaussian mixture models (with number of states = 4). Comparison of classification accuracies obtained using one-pass and two-pass algorithms

Two interest-point detectors, namely scale-invariant-feature-transform (SIFT) and

speedup-robust-features (SURF), have been tested. Both SIFT and SURF are local feature detectors and descriptors, which are used to detect interest points in an image and are robust to different image transformations. For our application, we needed a feature detector, which would be as robust to scale and angle changes as possible. During our experiments, it has been found out that SIFT-based features outperform the results obtained using SURF-based features. Various values of codebook size (k) have been used in k-means clustering of the interest points. We tested the accuracy with codebook size from 28 to 38, and the best results have been obtained with the value of a codebook size equal to 35. A plot depicting the comparison of results of classifica-

Fig. 8 Comparison of classification accuracy obtained using SIFT and SURF features. Accuracy variation with respect to codebook size

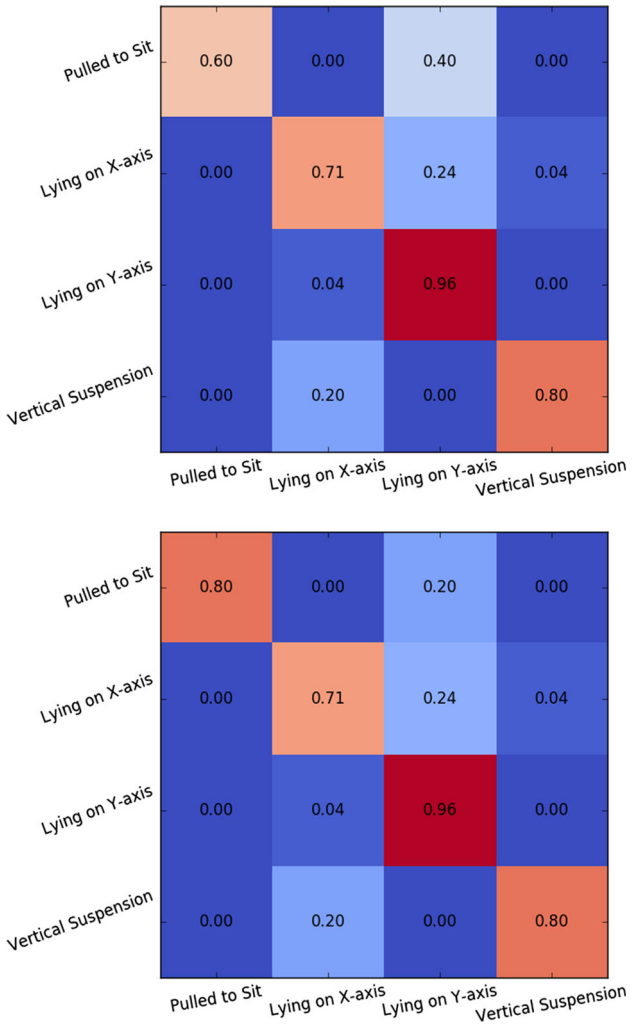

Fig. 9 Confusion matrices obtained without and with two-pass algorithm

Fig. 10 Frames from 3 different correctly classified sequences for each of the four classes, **a** pulled-to-sit, **b** laying-horizontally, **c** laying-vertically, and **d** vertical suspension

tion using SIFT- and SURF-based features, and variation of accuracy with codebook size, is presented in Fig. 8.

An accuracy of 79% has been obtained when classification is performed on four classes with maximum confusion being in *pulled-to-sit* and *laying-vertically* classes. These two classes are then merged, and a second pass of classification between the two classes has been performed using the features from infant's blob bounding box as mentioned in Sect. 3.1.

The confusion matrices without and with the two-pass algorithm, respectively, are shown in Fig. 9. It is evident from the first confusion matrix shown in Fig. 9 that *pulled-to-sit* class is misclassified as *laying-vertically* in more than 40% of the cases. After using the two-pass algorithm, however, the confusion between these two classes is eliminated to a great extent as evident in the second confusion matrix given in Fig. 9. The accuracy of classification for *pulled-to-sit* class improves by 20% after this second pass. This improves the overall accuracy to 84%. Some frames from correctly classified sequences (three each for the four different classes) are shown in Fig. 10.

4.3 Comparison with other classifiers

Recurrent neural networks (RNNs) can remember previous inputs and let them influence the network output. The RNNs, however, are only able to learn tasks that require a short-term memory and are thereby less effective when processing long sequences. The long short-term memory (LSTM) architecture provides remedies for the above-stated problem. LSTM networks have been tested in various applications like music improvisation, sequence labeling [27], and phoneme classification [28]. We have used a neural network architecture with one hidden LSTM layer. The input layer has size equal to

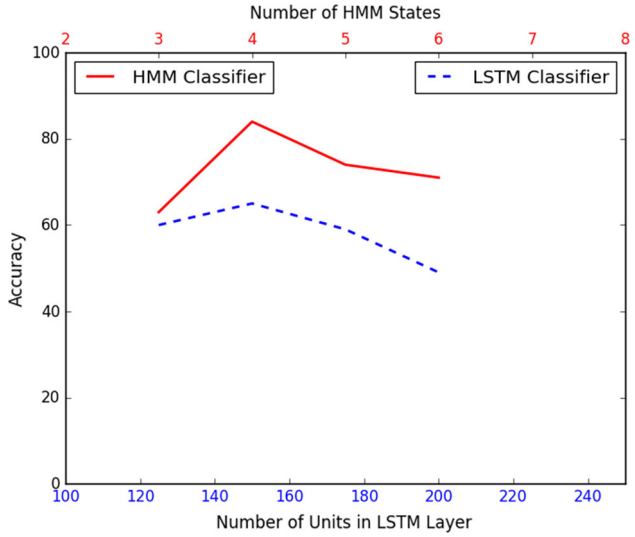

Fig. 11 Variation of classification accuracy with respect to number of units in LSTM layer and comparison of classification accuracy obtained using HMM and LSTM classifiers

the number of words in the codebook. For the output layer, the softmax nonlinearity has been used for 1 out of K classification tasks. The softmax function outputs the probability of a sequence belonging to each one of the K classes, and the sum of all probabilities is equal to 1. The hidden layer contains several unidirectional LSTM neurons that are fully interconnected and fully connected to the rest of the network. A LSTM layer with 150 units produces the best results for this classification task. The learning rate and the momentum have been set to 0.01 and 0.9, respectively.

Classification using LSTM has been performed on 5 videos using a network trained with 65 videos and with 10 iterations in the Monte Carlo cross-validation scheme. The LSTM layer with 150 units has been found to be the best for classification. The number of units less than 150 leads to noticeable divergence, while more than 150 leads to overfitting. An accuracy of 65% has been obtained with 150 units in the LSTM layer. A curve representing the variation of classification accuracy with respect to the number of units in LSTM layer is shown in Fig. 11.

4.4 Comparative analysis of various features

We also present a comparative analysis of results obtained using various popular features. As described in our previous work [16], frames are first converted to YC_rC_b color space and the skin color regions were detected. The image is then converted to a binary image based on the results of skin color segmentation. Morphological erosion using a 3×3 rectangular structuring element is performed on the binary image, which removes small spurious regions. A 5×5 morphological closing operation is then performed on the resulting binary image to make the infant's body's area more prominent. The largest blob (area-wise), which represents the infant's body, is chosen for further processing. This blob is then skeletonized using the thinning algorithm as described by Guo and Hall [29]. The results of segmentation upto the detection of the largest blob are shown in Fig. 5. The results of skeletonization are shown in Fig. 12 where (a) shows the binary image obtained after skin color segmentation and morphological operations (erosion and closing operations as described above) and (b) shows the results of skeletonization performed on the binary image using Guo and Hall's [29] thinning algorithm. Next, we have extracted skeleton

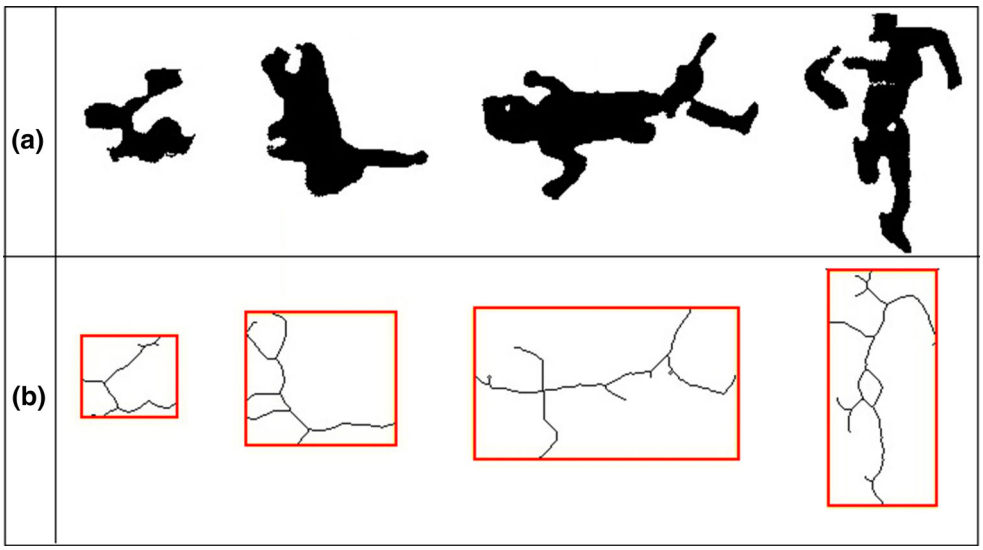

Fig. 12 Results of skeletonization. **a** Binary images after skin color segmentation, binarization, and morphological operations. **b** Binary images after skeletonization

Table 1 Comparison of accuracy obtained using different methods

Method	Accuracy (%)
Skeleton features + HMM [16]	64
Bag-of-words features + LSTM-RNN	65
Bag-of-words features + HMM (one-pass)	79
Bag-of-words features + HMM (two-pass)	84

features, namely width of skeleton bounding box, height of skeleton bounding box, aspect ratio of skeleton bounding box, distance between farthest nodes in the skeleton, center of mass of skeleton from the thinned images. We have performed classification using HMM.

In our proposed method using bag-of-words features, we have recorded accuracy of 65% using LSTM-RNN and 79% using HMM with the one-pass algorithm. To remove the confusion between *pulled-to-sit* and *laying-vertically* classes, we have used the aforementioned two-pass approach. We have merged the *pulled-to-sit* and *laying-vertically* classes into a single class and performed the first pass of classification using HMM. Then, each detected sequence of this combined class is checked during the second pass of the classification algorithm using HMM to assert whether it belonged to *pulled-to-sit* or *laying-vertically*. The overall accuracy improves to 84% after this two-pass classification. A comparison of these approaches is listed in Table 1.

4.5 Results of event segmentation

Experiments have been carried out on videos made by randomly joining 2–3 videos of individual exercises. The sum of magnitudes of optical flow velocities for each frame of the video sequence is calculated. The points with an abrupt increase in the magnitude of optical flow are identified as the position of an event change. A plot of magnitude of optical flow with respect to frame number of a video sequence is shown in Fig. 13. Figure 14 shows the frames where the spikes are observed as depicted in Fig. 13a, b. It is evident from the figure that the points of abrupt spikes are indeed the points where a change in the exercise occurred.

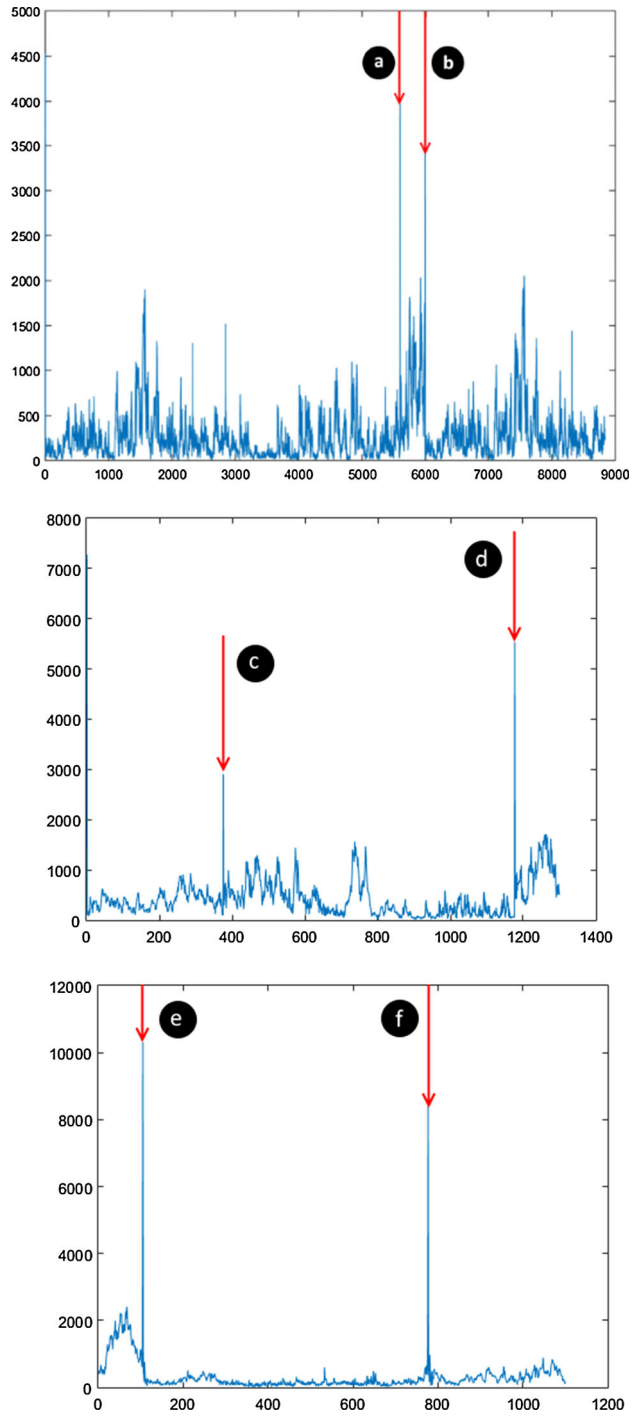

Fig. 13 Variation of optical flow velocity with respect to frame number. Frames corresponding to the potion of the spikes in time domain are shown in Fig. 14

5 Error analysis

The dataset is challenging because of the presence of objects which are not of our interest within the field of view of the camera. The presence of such objects often leads to wrong detection of local points using SIFT that are not points of interest. For instance, in some frames the expert's hands occupy a significant portion of the scene leading to occlusion and false detection of the infant's body. Some videos contain other objects such as toys used to pacify the crying babies. They also lead to detection of false interest points.

In case of event segmentation, any sudden movement may result in a spike causing noticeable change in the magnitude of optical flow. This results in wrong detection of the event

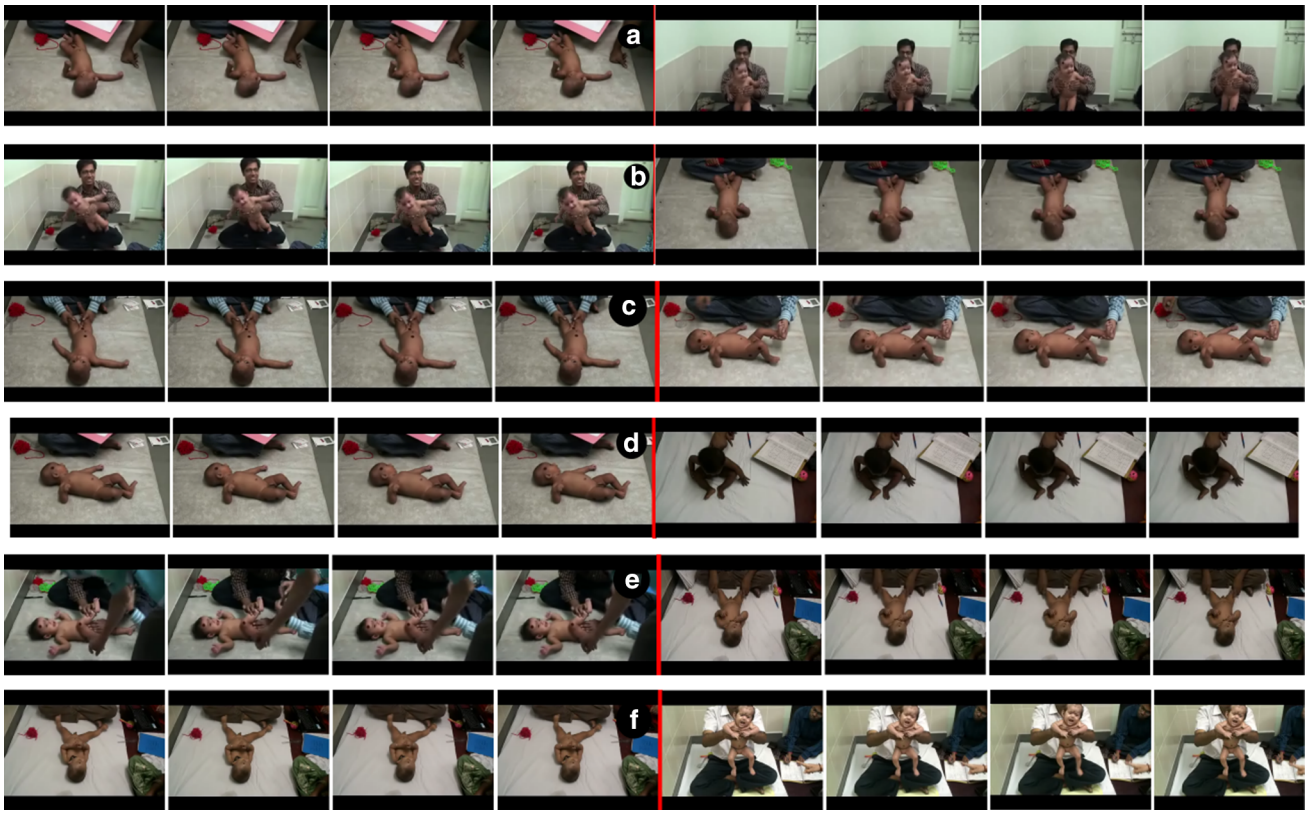

Fig. 14 Image frames depicting the spikes a–f as depicted in Fig. 13

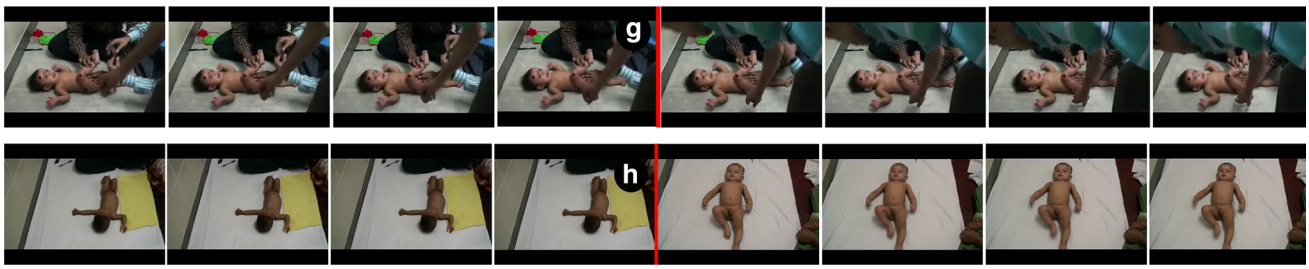

Fig. 15 Errors in event segmentation

points. This usually happens when the spikes are smaller in magnitude. As shown in Fig. 15, event segmentation at (g) happened because of sudden movement. In case of (h), wrong segmentation happened because of the sudden change even though there was no change in the original exercise class. However, this is an insertion error that essentially leads to detection of additional temporal points representing change events. This is not as bad as deletion errors. Insertion errors can be removed by tweaking the threshold parameter while looking for peaks optical flow magnitude. However, a significantly high value may result in deletion errors, wherein a change in exercise may remain undetected. A threshold value that is four times the average magnitude of optical flow across all frames was found to be suitable.

6 Conclusion

In this paper, we present an approach to classify exercises in the HINE videos using bag-of-words features. We have extracted SIFT features from every frame of a test video, and we have performed k-means clustering on the SIFT points to generate a codebook of visual words. Then, we have calculated the frequency of each visual word across all frames. Next, we have employed a two-pass classification algorithm by merging *pulled-to-sit* and *laying-vertically* classes during the first pass, and finally we have used infant's blob features to distinguish between these two classes in the second pass. We have observed that HMM-based classification performs better as compared to LSTM-RNN-based classification. We have obtained classification accuracy of 84% using HMM-based classifier. It has been observed that bag-of-words

features significantly outperform skeleton bounding box features by a margin of 19%, which is significant large. We have also presented a method of event segmentation in long videos of HINE test by joining videos that contain different exercises. We have calculated the magnitude of optical flow velocities for each frame and then looked for large and abrupt changes in the magnitude to detect temporal points where there is a change in exercise.

The proposed method has several applications. It can be used to preprocess HINE video recordings. Since the video recording of a single patient is usually done continuously, separating each exercise is very important. These segmented videos can be used for automatic or semiautomatic analysis as proposed by various researchers.

References

1. Zhang, R., Vogler, C., Metaxas, D.: Human gait recognition at sagittal plane. Image Vis. Comput. **25**(3), 321–330 (2007)
2. Liu, Q., Sclabassi, R., Sun, M.: Change detection in epilepsy monitoring video based on Markov random field theory. In: Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems, pp. 63–66 (2004)
3. Viola, P., Jones., M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: Ninth IEEE International Conference on Computer Vision, pp. 734–741 (2003)
4. Singh, S., Hsiao, H.: Infant telemonitoring system. In: Proceedings of the 25th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, vol. 2, pp. 1354–1357 (2003)
5. Nishida, Y., Motomura, Y., Kitamura, K., Mizoguchi, H.: Infant behavior simulation based on an environmental model and a developmental behavior model. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 1555–1560 (2004)
6. Dubowitz, L., Dubowitz, V., Mercuri, E.: The Neurological Assessment of the Preterm and Full Term Infant, Clinics in Developmental Medicine. Heinemann, London (2000)
7. Romeo, D., Guzzetta, A., Scoto, M., Cioni, M., Patusi, P., Mazzone, D., Romeo, M.: Early neurologic assessment in preterm infants: integration of traditional neurologic examination and observation of general movements. Eur. J. Paediatr. Neurol. **12**(3), 183–189 (2008)
8. Dogra, D., Majumdar, A., Sural, S., Mukherjee, J., Mukherjee, S., Singh, A.: Analysis of adductors angle measurement in hammersmith infant neurological examinations using mean shift segmentation and feature point based object tracking. Comput. Biol. Med. **42**(9), 925–934 (2012)
9. Dogra, D.P., Badri, V., Majumdar, A.K., Sural, S., Mukherjee, J., Mukherjee, S., Singh, A.: Video analysis of Hammersmith lateral tilting examination using Kalman filter guided multi-path tracking. Med. Biol. Eng. Comput. **52**(9), 759–772 (2014)
10. Dogra, D.P., Majumdar, A.K., Sural, S., Mukherjee, J., Mukherjee, S., Singh, A.: Automatic adductors angle measurement for neurological assessment of post-neonatal infants during follow up. Pattern Recognit. Mach. Intell. Lect. Notes Comput. Sci. **6744**, 160–166 (2011)
11. Dogra, D.P., Majumdar, A.K., Sural, S., Mukherjee, J., Mukherjee, S., Singh, A.: Toward automating Hammersmith pulled-to-sit examination of infants using feature point based video object tracking. IEEE Trans. Neural Syst. Rehabil. Eng. **20**(1), 38–47 (2012)
12. Romeo, D., Cioni, M., Scoto, M., Mazzone, L., Palermo, F., Romeo, M.: Neuromotor development in infants with cerebral palsy investigated by the Hammersmith Infant Neurological Examination during the first year of age. Eur. J. Paediatr. Neurol. **12**(1), 24–31 (2008)
13. Dey, P., Dogra, D.P., Roy, P.P., Bhaskar, H.: Autonomous vision-guided approach for the analysis and grading of vertical suspension tests during HINE. In: Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 863–866 (2016)
14. Roy, S., Dogra, D.P., Bhattacharya, S., Saha, B., Biswas, A., Majumdar, A.K., Mukhopadhyay, J., Majumdar, B., Singh, A., Paria, A., Mukherjee, S.: A web enabled health information system for Neonatal Intensive Care Unit (NICU). In: Proceedings of the 7th IEEE World Congress on Services (SERVICES), pp. 451–458 (2011)
15. Dogra, D.P., Nandam, K., Majumdar, A.K., Sural, S., Mukhopadhyay, J., Majumdar, B., Mukherjee, S., Singh, A.: A tool for automatic Hammersmith infant neurological examination. E-Health Med. Commun. **2**(2), 1–13 (2011)
16. Ansari, A.F., Roy, P.P., Dogra, D.P.: Posture recognition in HINE exercises. In: Proceedings of International Conference on Computer Vision and Image Processing (CVIP) (2016)
17. Poleg, Y., Arora, C., Peleg, S.: Temporal segmentation of egocentric videos. In: CVPR, pp. 2537–2544 (2014)
18. Ballan, L., Bertini, M., Bimbo, A.D., Serra, G.: Action categorization in soccer videos using string kernels. In: Seventh International Workshop on Content-Based Multimedia Indexing, pp. 13–18 (2009)
19. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
20. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM J. Res. Dev. **1**(4), 315 (1957)
21. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **28**, 11–21 (1972)
22. Chai, J., Ngan, K.: Face segmentation using skin-color map in videophone applications. IEEE Trans. Circuits Syst. Video Technol. **9**(4), 551–564 (1999)
23. Hu, J., Lim, S.G., Brown, M.K.: Writer independent on-line handwriting recognition using an HMM approach. Pattern Recogn. **33**(1), 133–147 (2000)
24. Huang, X.D., Arikiand, Y., Jack, M.A.: Hidden Markov Models for Speech Recognition. Edinburgh University Press, Edinburgh (1990)
25. Yang, J., Xu, Y.: Hidden Markov model for gesture recognition. Technical Report, Robotics Institute, Carnegie Mellon University, 10 (1994)
26. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. Int. J. Comput. Vision. **12**(1), 43–77 (1994)
27. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Action classification in soccer videos with long short-term memory recurrent neural networks. In: 20th International Conference on Artificial Neural Networks, pp. 154–159 (2010)
28. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks. **18**(5), 602–610 (2005)
29. Guo, Z., Hall, R.W.: Parallel thinning with two-subiteration algorithms. Commun. ACM **32**(3), 359–373 (1989)

Abdul Fatir Ansari is a B.Tech. student at IIT Roorkee. His area of interest is computer vision and its applications in medical examinations.

Dr. Partha Pratim Roy obtained Ph.D. degree from Autonomous University of Barcelona, Spain, in 2010, MS degree from Autonomous University of Barcelona, Spain, and B.Tech. from University of Kalyani, India, in 2002. Presently, he is an Assistant Professor at IIT Roorkee, India. He was with Samsung Research Institute, India (2013-2014). He worked as Postdoctoral Research Fellow in Synchromedia Lab, Canada, and RFAI Lab (2010-2013), France. He has worked as Assistant System Engineer in TCS, India (2003-2005). Dr. Roy has published more than 60 research papers. His research interests include pattern recognition, multilingual text recognition, biometrics, and computer vision.

Dr. Debi Prosad Dogra obtained Ph.D. degree from IIT Kharagpur, India, in 2012, M.Tech. degree from IIT Kanpur, India, in 2003, and B.Tech. from Haldia Institute of Technology, India, in 2001. Presently, he is an Assistant Professor in the School of Electrical Sciences, IIT Bhubaneswar, India. He was with Samsung Research India (2011-2013). He worked with ETRI, South Korea (2006-2007). He worked as a faculty at HIT (2003-2006). He has published more than 50 research papers in international journals and conferences. His research interests include visual surveillance, augmented reality, and human computer interface. He has obtained three US patents.