Infant Action Database: A Benchmark for Infant Action Recognition in Uncontrolled condition

S Sujitha Balasathiya, S Mohamed Mansoor Roomi, B Sathyabama

M.E, ECE, Thiagarajar College of Engineering, Madurai, India.
Professor, ECE, Thiagarajar College of Engineering, Madurai, India.
Associate Professor, ECE, Thiagarajar College of Engineering, Madurai, India.

Corresponding author’s e-mail: sujithabalasathiya21@gmail.com

Abstract: The main focus of our work is to create a database for action recognition of unattended infants in uncontrolled environment with wide variations in surroundings, lighting, interactions with objects, camera motions, etc., Action recognition of infants is emerging as an important and technically challenging computer vision problem as compared with adult action recognition because of their physical appearance. Most of the previous action recognition techniques have focused on the recognition of action captured under a controlled environment in a standard laboratory setting. In this study a novel database is introduced that can be used as a benchmark for surveillance parenting. This database involves nine normal and nine abnormal actions classes which consist of actions and movements of infants occurring in fairly uncontrolled conditions. This database consists of realistic user-uploaded videos which are recorded in the clustered background and different camera motion. After the collection of all videos, they are manually trimmed to form a database. To further evaluate the performance of the database, HOG features were extracted from database set and trained by different Machine Learning classifiers like Multi-class Naive Bayes, Support Vector Machine, Ensemble classifier, Discriminant analysis and Decision tree classifiers. This experimentation shows that the database is complex and robust that can serve as a base for testing action recognition algorithms.

1.Introduction

The majority of existing action recognition database is mainly designed for adult actions e.g., HMDB51 [10], UCF101 [9] and some of the databases contains videos that are recorded in the unrealistic environment e.g., KTH [4], UCF Sports [11]. For instance, they contain movie clippings and videos that are staged by professional actors. We addressed both these issues in this project.

Prerequisite to this database creation is, one should know, up to which age a child is an infant. With the literature’s available online sources and governmental sources, we concluded that up to 3 years child may be called an infant. Recently, new databases have been created for Gender classification [13], Banana ripeness classification [12], INFACE database [15] for variation in expression and poses of an individual and EPOCH database [15] for variation in age of an individual, Roshan Singh conducted a survey and produced benchmark datasets for classification in each area, as well as a collection of characteristics by which datasets could be compared [14]. As a result, we developed a new database for infant actions recognition by referencing all of these databases. Generally, infant actions can be
classified into two types Normal and Abnormal actions. We can use this database to take care of unattended infants’ home, creche and in a separate room.

Most of us live in a nuclear family (parents and children) and almost in all the families both the parents are working. So, they cannot spend quality time with the infant but generally, the infant needs 24*7 attention from the parents. To overcome this situation, we can train a machine using Artificial Intelligence that can imitate intelligent human behaviour. More specifically we can train the machine using deep learning to imitate the intelligent human brain. For this kind of model, we can use this infant action recognition database.

2. Database Details
Actions that will not risk an infant's life are categorized under Normal actions of the infant-like Crawling, Crying, Laughing, Sleeping, Standing, Walking, Playing, Sitting, Eating.

![Figure 1: Samples of Infant Action - Normal](image1)

Actions of an infant that leads to self-injury, suffocation, loss of life, damage to articles/facilities, injury to others are categorized into Abnormal actions of infant-like Infant escaping from crib or gate, Infant in steps, Infant falling down, Infant eating sand, Infants with pet animals, Infant in water, Infants fighting, Infant with a knife.

![Figure 2: Samples of Infant Action - Abnormal](image2)
Figure 3. The number of clips per abnormal action class of infant and distribution of clip lengths is illustrated by the colours.

Figure 4. The number of clips per normal action class of infant and distribution of clip lengths is illustrated by the colours.

**Clip Group:** The clips of each action are sub-categorized into several groups based on their environment, interactions with the objects and so on. Each group contains around 15-20 clips. The bar chart of figure 3 and figure 4 shows the no. of clips per normal and abnormal actions of the infant. The distribution of clip lengths is illustrated by the colours. Once an abnormal action of the infant is detected, immediate action has to be taken by the protectors to rescue the infant. So, this database consists of short duration videos using which we can recognise the action immediately.
Figure 5. The total time of videos for each class is illustrated using the blue bars. The average length of the clips for each action is depicted in orange.

The videos for this database are downloaded from online sources like YouTube, Instagram and inappropriate contents are manually trimmed for each action. For each clip, audio is preserved in this database. Table 1 illustrates the features of the database. Generally, abnormal actions of infants are of two types actions that risk their life like Infant falling down, Infant with a knife, Infant in water and actions that cause injuries and health problems like Infant escaping from crib or gate, Infant in steps, Infant eating sand, Infants with pet animals, Infants fighting. Our database contains 1095 clips with an overall duration of 36.75 minutes. The minimum and maximum length of clips is 1 sec and 11 secs respective. For the infant actions and each class contains around 1-4 groups based on their environment.

Table 1: Illustrates the characteristics of the database

| Actions                | 18  |
|------------------------|-----|
| Clips                  | 1181|
| Groups per Action      | 1- 4|
| Clips per Group        | 40-120|
| Mean clip length       | 1.80 sec|
| Total duration         | 40.5 min|
| Min clip Length        | 1 sec|
| Max clip Length        | 11sec|
| Frames Rate            | 26 Fps|
| Audio                  | Yes |

The videos for this database are downloaded from online sources like YouTube, Instagram and inappropriate contents are manually trimmed for each action. For each clip, audio is preserved in this database. Generally, abnormal actions of infants are of two types actions that risk their life like Infant falling down, Infant with a knife, Infant in water and actions that cause injuries and health problems like Infant escaping from crib or gate, Infant in steps, Infant eating sand, Infants with pet animals, Infants fighting. Our database contains 1095 clips with an overall duration of 36.75 minutes. The minimum and maximum length of clips is 1 sec and 11 secs respective. For the infant actions and each class contains around 1-4 groups based on their environment.
3. Pre-processing

Pre-processing is one of the basic and most important steps in computer vision problems. By using only, the appearance features from RGB frames we cannot effectively differentiate the classes like Fast walking and Running because they are having Suttle difference between them. To overcome this, we have also extracted motion features along with the spatial features. For spatial features, videos are converted into frames and keyframes are extracted from frames and for temporal features, the optical flow of the video is computation using the Lucas and Kanade method [5] which computes the speed and direction of movement of each pixel between each frame.

3.1 Lucas and Kanade method

In this method, the motion information u and v are calculated separately for each and every pixel of a video with an assumption that pixels in the locality have same u and v i.e., if we consider a pixel (x, y) for optical flow calculation all its neighbor pixel will have same motion information’s. [1]

\[
I_x t u + I_y t v = -I_t
\]

\[
[I_x I_y] [u v] = -I_t \tag{1}
\]

\[
A x = b
\]

\[
I_x, I_y, I_t \tag{2}
\]

Values can be calculated by convolving respective masks with frames at time t and t+dt then compute average of both the values.

\[
I_x = 0.5 \times (\text{Frame t} \ast I_{x1} + \text{Frame t+d} \ast I_{x2}) \tag{3}
\]

\[
I_y = 0.5 \times (\text{Frame t} \ast I_{y1} + \text{Frame t+d} \ast I_{y2}) \tag{4}
\]

\[
I_t = (\text{Frame t} \ast I_{t1} + \text{Frame t+d} \ast I_{t2}) \tag{5}
\]

Compute the determinant value of matrix A [2] if it is zero skip the calculation of motion information u and v because the pixel represents a stationary object. If it is non-zero then calculate the values using Least Square Method [6].

\[
x = A^{-1} b
\]

\[
A^T A x = A^T b
\]

\[
x = (A^T A)^{-1} A^T b \tag{6}
\]

4. HOG Feature Extraction

Histogram of Oriented Gradient (HOG) is one of the most common feature descriptors used in Computer Vision problems to focus on shape or structure of an object along with edge direction by extracting gradient and orientation of the edges. For each class, the dataset is divided into training set (70%) and test set (30%). Then pre-process the width and height ratio of the image as 1:2 to divide the image into 32x32 patches for feature extraction. Gradient magnitude and direction of each pixel of an image has to be calculated by computing the difference in pixel intensities along X (G_x) and Y (G_y) direction which is similar to using Sobel Kernal of size 1. As a result, two new matrices will be generated which is used to determine the Gradient Magnitude [7] and Orientation [8] for each and every pixel.

\[
\text{Total Gradient Magnitude} = \sqrt{G_x^2 + G_y^2} \tag{7}
\]

\[
\tan(\Phi) = \frac{G_x}{G_y}
\]

\[
\Phi = a \tan \left( \frac{G_x}{G_y} \right) \tag{8}
\]

Using this gradient and orientation(direction) histogram is created with variables in x-axis and frequency in y-axis for single cell of size 32x32. To avoid lighting variation normalisation has to be done for each cell and then concatenated to form a block. For each and every image in the training and testing set HOG features matrix and label was extracted.
5. Evaluation of database

In order to evaluate the efficiency of the proposed database, the learnt feature was further used to train some Machine Learning algorithms such as Multi-class Naïve Bayes, Support Vector Machine, Ensemble classifier, Discriminant analysis and Decision tree classifiers. Finally achieved a mean classification accuracy of 30%, 71%, 27%, 66% and 30% respectively shown in figure 6. So, the average classification accuracy for this database is 44.8. As a result, we have proved the complexity of the infant action database.

![Classification accuracy graph](image)

**Figure 6:** Comparison of different ML classifiers

6. Conclusion

To the best of our knowledge, this was the first-ever database created for infant actions which consist of an acceptable number of classes, a large number of clips and realistic user-uploaded videos which are recorded in unconstrained condition with incorporate camera movement, different lighting conditions which can be used for surveillance parenting. It consists of 18 actions classes for normal and abnormal actions of an infant which are downloaded from YouTube and Instagram. As a future work we can use this database to create a surveillance parenting model by monitoring the activities of unattended infants. This model is mainly designed for unattended infants and using this model we cannot control the event from it happening but we can alert the protectors to pay attention to the infant once the event has happened because sometimes lack of attention may lead to death.

Reference

[1] Baykal, E., Dogan, H., Ercin, M.E. et al. Transfer learning with pre-trained deep convolutional neural networks for serous cell classification. Multimed Tools Appl 79, 15593–15611 (2020). https://doi.org/10.1007/s11042-019-07821-9

[2] C. Li, A. Pourtaherian, L. Van Onzenoort, W. E. Tjon a Ten and P. H. N. De With, "Infant Facial Expression Analysis: Towards A Real-time Video Monitoring System Using R-CNN and HMM," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.3037031.

[3] C. Li, A. Pourtaherian, W. E. Tjon a Ten and P. H. N. De With, "Infant Monitoring System for Real-Time and Remote Discomfort Detection," 2020 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2020, pp. 1-2, doi: 10.1109/ICCE46568.2020.9043065.

[4] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach, 2004. International Conference on Pattern Recognition (ICPR).
[5] C. Chang, C. Chou and E. Y. Chang, "CLKN: Cascaded Lucas-Kanade Networks for Image Alignment," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3777-3785, doi: 10.1109/CVPR.2017.402.

[6] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249-256. 2010.

[7] H. Zhang, D. Liu and Z. Xiong, "Convolutional Neural Network-Based Video Super-Resolution for Action Recognition," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, 2018, pp. 746-750, doi: 10.1109/FG.2018.00117.

[8] J. Kim and C. S. Won, "Action Recognition in Videos Using Pre-Trained 2D Convolutional Neural Networks," in IEEE Access, vol. 8, pp. 60179-60188, 2020, doi: 10.1109/ACCESS.2020.2983427.

[9] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes from Videos in The Wild, CRCV-TR-12-01, November, 2012.

[10] Kuehne, Hilde & Jhuang, Hueihan & Garrote, Estibaliz & Poggio, Tomaso & Serre, Thomas. (2011). HMDB51: A Large Video Database for Human Motion Recognition. Proceedings of the IEEE International Conference on Computer Vision. 2556-2563. 10.1109/ICCV.2011.6126543

[11] M. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatiotemporal maximum average correlation height iter for action recognition, 2008. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[12] M. Senthilarasi, S. Md. Mansoor Roomi, “Particle Swarm Optimized Fuzzy model for classification of banana ripeness”, IEEE Sensors, Vol. 17, No. 15, 2017.

[13] M. Annalakshmi,S.Md.Mansoor Roomi,A.Sheik Naveedh, “A hybrid technique for gender classification with SLBP and HOG features”, Springer, Cluster computing, pp 1–10, 2018.

[14] Singh, R., Sonawane, A. & Srivastava, R. Recent evolution of modern datasets for human activity recognition: a deep survey. Multimedia Systems 26, 83–106 (2020). https://doi.org/10.1007/s00530-019-00635-7.

[15] S.Md.Mansoor Roomi, M.Parisa beham, " INFACE and EPOCH database: A Benchmark for Face Recognition in Uncontrolled conditions", Research Journal of Applied Sciences, Engineering and Technology, Vol.8 Issue.6, Pg.no: 795-802, 2014.

[16] Y. Lee, K. K. Kim and J. H. Kim, "Prevention of Safety Accidents through Artificial Intelligence Monitoring of Infants in the Home Environment," 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea (South), 2019, pp. 474-477, doi: 10.1109/ICTC46691.2019.8939675