Journal of Computer Science 10 (2): 296-304, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.296.304 Published Online 10 (2) 2014 (http://www.thescipub.com/jcs.toc)

CHILD VIDEO DATASET TOOL TO DEVELOP OBJECT TRACKING SIMULATES BABYSITTER VISION ROBOT

[1]Hanan Aljuaid and [2]Dzulkifli Mohamad

[1]Faculty of Computer Science and Info Systems, Taif University, Taif, Saudi Arabia
[2]Faculty of Computer Science and Info Systems, University Technology Malaysia, Johor Bharu, Malaysia

Received 2013-09-04; Revised 2013-09-06; Accepted 2013-11-16

ABSTRACT

This study presents a Child Video Dataset (CVDS) that has numerous videos of different ages and situation of children. To simulate a babysitter's vision, our application was developed to track objects in a scene with the main goal of creating a reliable and operative moving child-object detection system. The aim of this study is to explore novel algorithms to track a child-object in an indoor and outdoor background video. It focuses on tracking a whole child-object while simultaneously tracking the body parts of that object to produce a positive system. This effort suggests an approach for labeling three body sections, i.e., the head, upper and lower sections and then for detecting a specific area within the three sections and tracking this section using a Gaussian Mixture Model (GMM) algorithm according to the labeling technique. The system is applied in three situations: Child-object walking, crawling and seated moving. During system experimentation, walking object tracking provided the best performance, achieving 91.932% for body-part tracking and 96.235% for whole-object tracking. Crawling object tracking achieved 90.832% for body-part tracking and 96.231% for whole object tracking. Finally, seated-moving-object tracking achieved 89.7% for body-part tracking and 93.4% for whole-object tracking.

Keywords: Object Detection and Tracking, Object Detection and Tracking Dataset, Body Part Tracking, Computer Vision, Robot Vision, Babysitter Robot Vision, GMM

1. INTRODUCTION

Body parts tracking from monocular a video sequence has been one of the research areas with an increasing number of technical applications. It is mainly solved in controlled situations where several calibrated cameras are used with babysitter robot vision in tracking babies and toddlers.

After detected the object, the tracker's task must perform to find its equivalence in the subsequent frames though constructing object's trace. In the image, which is conquered with the object tracking may also provide the complete region of the object.

The tasks of founding matching between the object illustrations thru frames could be either implemented jointly or separately. In the jointly case, updating object location information to detect the object region that is jointly estimated and gotten from preceding frames. In the separately case, processes of an object detection and tracking algorithm occupy object regions across frames (Vedaldi and Soatto, 2006).

Major challenges presented in robust and accurate tracking of non-rigid and object that moving fast deprived of reaching controlled to certain model assumptions. Some can streamline tracking by impressive limitations on motion of objects. For instance, nearly all-tracking algorithms accept the object motion to remain smooth without sudden variations. Other constrains the object motion to occur with constant acceleration, or constant rapidity centered upon an initial information, such as the size and the number of objects, or the object presence and shape that simplify the problem.

Corresponding Author: Hanan Aljuaid, Faculty of Computer Science and Info Systems, Taif University, Taif, Saudi Arabia

For business applications, they should perform mostly partitioned indoors spaces where sudden illumination changes, for instance due to on-and-off of a light switch, can occur. They are expected to handle significant object size changes due to oblique views and severe occlusions due to usually lower camera heights. They need to resolve multiple object tracking even often the descriptors are insufficient as people tend to dress in for instance dark clothes in business environments (Stenger *et al*., 2001; Rosales and Sclaroff, 1999).

However, there are a real time tracking people method and their body parts in homochromatic images (Haritaoglu *et al*., 1998). Haritaoglu *et al*. (1998) concepts dynamic models of object's movements called 4W. It theories object tracking models that can track persons occlusion events in the images. The W models were proposed to control categories of actions. Correspondingly, it controlled to tracking persons who are absorbed in detecting from a affecting observation plat.

In this approach, the tracking models occupied for babysitter vision to track babies and toddlers and their parts according to the moving situation. This method start by libelling the head, upper and lower sections in three situation walking, crawling and seated moving in order to track the action. A novel GMM algorithm applying in each libelling area to track the move of body parts in this area (Shi and Malik, 2000; Sasaki *et al*., 2009; Torralba *et al*., 2007). To avoid the error of ambiguous object's shape in some frame and object depth in the camera a texture template is used at each previse frame (Kanhere *et al*., 2005; Sarfraz *et al*., 2011; Aljuaid *et al*., 2010; 2009).

1.1. Child's Object Video Dataset (CVBS)

The data collected in this dataset are designed to be realistic, natural and challenging for video surveillance researches in terms of its variety in scenes, resolution, background clutter and children activity or event categories. The following sections discuss the hardware that used to built CVBS and it is includes.

1.2. Hardware

Using a single camera with the viewing plane perpendicular to the ground plane, an outdoor and indoor-space at two view angles: a 45° path (angle-view) toward the camera and a frontal-parallel path (side-view) in relation to the viewing plane of the camera. The side-view data was captured at two different depths, 3.9 and 8.3 m from camera.

Sequentially, Video data was captured using HDR-PJ790E Camcorder with 26 mm Wide-angle Lens. It is maximum still image resolution 24.1 megapixels with 3984×2240 Resolution. The image sensor for the camcorder is seExmor CMOS Sensor. It has the built-in LED video light and Electronic View Finder (Evf). The EVF provides a clear and crisp representation of the videos and allow to frame the videos better and increase stability to shoot the footage with precision. The camcorder allows capturing HD Video Codec AVCHD format ver.2.0 compatible: MPEG4- AVC/H.264. STD Video Codec, MPEG2-PS: MPEG-2 (Video) MP4: MPEG-4 A VC/H.264.

1.3. Dataset Encompasses

The data collected in this dataset are different size of babies and toddlers video of two types: Indoor babies and toddlers videos and outdoor babies and toddlers videos. Indoors baby and toddler videos are video's frames of baby or toddler object walking or crawling in close environment; this type of date, which needed in this research. In contrast, outdoor babies and toddlers videos are video's frames of baby or toddler object walking or crawling in open environment (Aljuaid *et al*., 2010). Videos was collected in natural scenes showing babies and toddlers performing normal actions with uncontrolled. There are normal attendant movers and background activities.

The contents of the indoor data are 100 videos and 55 videos for outdoor with different background. Many videos run across a varied range of spatial and temporal resolutions. The dataset affords the original videos with HD quality.

The videos were classified into eight groups according to the baby and toddler age, baby and toddler position and the environment as shown in **Table 1**. The babies and toddlers ages ranged from 4 months to about 6 years. The videos belonging to group 1 were babies from 4 to 8 months in the different position of baby in these ages such as sit down, creeping and crawling. Babies and toddlers of different age and position groups were given dataset matched their age and position level.

The videos collected sorted in a dataset. The dataset has double directories. One has 45 video of outdoor background and the other has 100 video of indoor background. Thus, the resulting database has 145 videos each video has more than 100 frames, as appear in the two **Table 2 and 3**.

As revealed, the dataset is has fixable samples with the different ages and position of altered babies and toddlers, which give the dataset more readability, as given away in **Table 3**.

Table 1. The dataset groups

Group	Childs age	Child position	Background
CIn	4-8 months	Creeping	Indoor-crowd
SCIn	4-8 months	Sit-down-crawling	Indoor-crowd
SCOut	8-36 months	Sit-down-crawling	Outdoor-crowd
SCIn2	8-36 months	Sit-down-crawling	Indoor-crowd
WCIn	3-5 years	Walking-creeping-crawling	Indoor- crowd
WCCOut	3-6 years	Walking-creeping-crawling	Outdoor-crowd

Table 2. Quantity of videos and frames in the dataset

Background	Video	Frame
Indoor	100	181840
Outdoor	55	13985

Table 3. Number of videos and frames in groups

		Database	
Child ages	Child's no	Video	Frames
4-8months	10	30	42151
8-36months	15	45	62420
3-4 years	12	43	62234
5-6 years	12	27	31020
Total	44	155	197825

1.4. Annotation

One of the challenge tasks in design a large dataset is annotating that need two major compromise factor, which is quality and cost. Especially, LabelMe designed a drawing tool for annotating static images. These tools developed to be compatible with CVDS.

The Moving objects (baby or toddler) in the CVDS are patent by bounding boxes, where the visible parts of it are labeled and they are not induced outside closure by guessing. For instance, if upper part of a baby is the visible part, then, only the upper part is labeled as object. This attention is essential to allow the detection and tracking algorithm to measure the performance of multiple moving objects more correctly.

The bounding boxes are labeled annotated by experts. It is should be as tight as possible and should not infer outside the objects actuality labelled and cover all related parts are captured in the bounding boxes. For instance, all the detectable parts of baby and toddler should be in the bounding box. Infrequently, additional important static parts that in the scenes such as games, adult person and other related objects are labeled when possible.

(A)

(B)

Fig. 1. Example of the notation in (A) one object walking the information of the object and the important static parts as door and Backpack appear when clicks the object in any frame as in frames 12 and 20. (B) Another example with two objects its information appears when clicks it in any frame

The annotation process is designed by clicking control points along the border of an object to form a bounding box. When the bounding box is adjusted, the notation process encouraged the user for the type of the object if he is baby or toddler, his gender girl or boy, his age, his location indoor or outdoor and information about his motion. The notation information is recorded and broadcast across all frames in the video while if the object were static or moving at all times during the sequence as shown in **Fig. 1**.

2. MATERIALS AND METHODS

The goal of the tracking method was to estimate the position of the object in each scene. However, the tracking algorithm had to continue tracking the object in low-level detection settings utilizing algorithms that depended on the split-merge-based region-growing method. This method can track the moving child object in three situations: Walking, crawling and seated moving depending on the child-body-part labeling technique that was applied to the object in the previous detection method. The next section discusses the body-part labeling technique and the tracking of these parts in each child-object.

2.1. Body-Part Labeling

Child's object is divided to body parts that are labelled by analyzing the contour of the object that detected by region growing model in each video frame. After a contour is produced, a bounding box is located on the contour area and separated into three parts: Head part, upper part and lower part.

There are three situations for the moving child's object: Walking, crawling and seated moving. However, in any situation the head is located through location the centroid of the pixels positioned in section. The upper part is located by finding the first and last of the pixels positioned in section the lower part lines the lower legs or legs and hands in the crawling situation as shows in **Fig. 2**. The lower section is subdivided into region 1 and 2. The distance (LR) within region 1 and 2, where the pixel place thru the maximum distance in every region is labeled region1 and region 2.

2.2. Tracking the Child's Body Parts

In addition to tracking the moving child-object as a whole, the algorithm was positioned to track body parts such as the head, hands, legs and feet in order to understand the action. As mentioned, there were three moving situations, i.e., walking, crawling and seated moving. Conversely, to track each moving situation, the algorithm employed a combination of the body-part measuring area by using GMM and template matching was used when an object was occluded and its shape could not be easily predicted for tracking. On the other hand, the body-part measurements could change according to the object depth in the image. To solve this, the algorithm measured the object depth at an angle facing towards the camera by utilizing the depth compensation method (Johnson and Bobick, 2001). In the next sections, the tracking method of the three moving situations using GMM will be discussed.

2.2.1. Main Features of the Tracking Method

The body-part labeling and the boundary box around the shape were utilized to measure the four main features of the tracking method, which were:

- H: The bounding box height around the silhouette
- L: The head area detected by the labeling point and the last labeling point in the upper section locations, as shown in **Fig. 3**
- D: The upper area selected by the first and last labeling points in the upper section
- LR: Region1, region2 and the horizontal distance between the left and right in the lower section

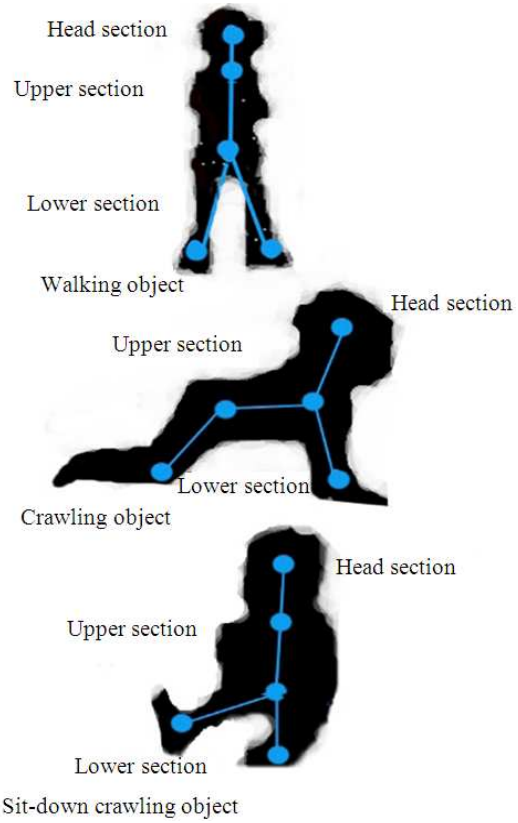

Fig. 2. The body silhouette regions in the walking, crawling and seated moving of the child-object

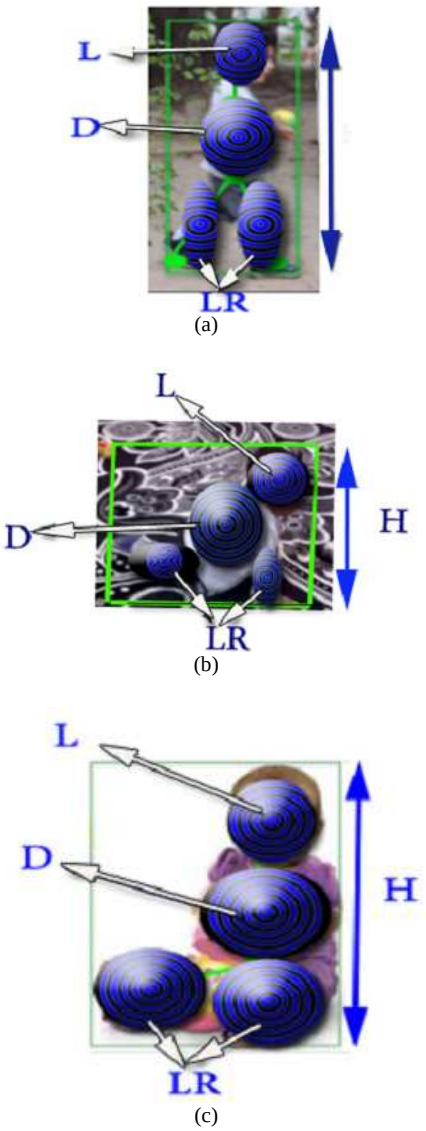

Fig. 3. The features areas f in the (a) walking object, (b) crawling object and (c) seated moving object

These features were sorted respectively in a moving tracking vector called the f vector, which comprised the following four measures:

$$f = [H, L, D, LR]$$

These measures updated continuously according to the labeling parts, texture template and depth compensation.

2.2.2. Moving-Tracking Alteration

The bounding box tracked the moving object continuously. On the other hand, the moving object had the potential to change its moving situation from one to another, which produced variations in how the body-part labeling techniques worked. The algorithms used GMM to calculate the Gaussian of each feature area in order to track the alterations in the moving object, as shown in **Fig. 3**:

$$P(x) = \sum_{i=1}^{k} w_{f,i,t} \times \eta\left(x, \mu_{f,i,t}, \Sigma_{f,i,t}\right)$$

where, $w_{f,i,t}$ is an estimate weight of kth Gaussian in the mixture of the feature area *f* at time t, $\eta(x, \mu_{f,i,t}, \Sigma_{f,i,t})$ is the Gaussian probability density function of the feature area f, $\mu_{f,i,t}$ is the mean value and $\Sigma_{f,i,t}$ is the covariance matrix. The first feature areas selected initialized the mean value of the k Gaussian distributions in each area:

$$\mu_f = \left[\mu_H, \mu_L, \mu_D, \mu_{LR}\right]$$

The variance could be created with an initial high value of each area:

$$\Sigma_f = \left[\Sigma_H, \Sigma_L, \Sigma_D, \Sigma_{LR}\right]$$

The weight value could be set for each feature area as:

$$w_f = \left[\frac{H}{k}, \frac{L}{k}, \frac{D}{k}, \frac{LR}{k}\right]$$

If the pixel in one of the feature area matched one of the k Gaussian distributions, the mean and variance were updated to adapt to the changes in the moving situation. The weight value of the k Gaussian distribution was updated in each feature area as follows:

$$w_{f,t} = (1-\alpha) w_{f,t-1} + \sigma Q$$

where, α is the learning rate and Q is 1 for the matched Gaussian of one feature area and 0 for the remaining Gaussian. The updated equations of μ and σ are as follows:

$$\mu_{f,i,t} = (1-\rho)\mu_{f,i,t-1} + \rho . X_{f,t}$$

$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + p\left(x_{f,t} - \mu_{f,t}\right)^T \left(X_{f,t} - \mu_{f,t}\right)$$

where, $\rho = \alpha\eta (X_f|\mu_{f,i}, \sigma_{f,i})$. If no match was found, the last distribution was replaced by a new Gaussian with the current values as its mean as well as an initially high variance and a low weight parameter.

3. RESULTS AND DISCUSSION

Our purpose in this study is to track a child-object and the moving body parts of the object. To achieve this aim, the head, upper and lower sections were labeled to apply the GMM to each section.

The tracking algorithm pursued several directions to improve its performance and to extend its capabilities. Firstly, the labeling used to image the body sections and positions was controlled by the child-object. The algorithm was able to recognize and track the child-object in three situations: Walking, crawling and seated moving. Secondly, we examined the algorithm within the CVDS dataset. This dataset had six groups of videos with children in the three situations, i.e., walking, crawling and seated moving. The dataset comprised 154 videos with 194,825 frames.

The examination was undertaken in three phases, i.e., an examination of the labeling of the body part sections, an examination of the detection of the area of each section and an examination of the performance of GMM to track the object and the body parts that applied in the CVDS dataset. **Figure 4** shows the performance of the algorithm in the three phases, where H denotes the performance of the tracking algorithm of the whole child-object. H provided a better performance than L, D and LR in the sections of body-part tracking across the three examination phases. Indeed, the examinations of whole-object tracking after applying GMM achieved 97.326%. On the other hand, our examination of the tracking of the head section (denoted by L) achieved 95.876% in the first phase, 91.243% in the second phase and 93.347% in the third phase. The performance of the tracking algorithm for the upper section D at 88.465% was lower than the performance of L and better than the Lower section (LR) performance. The performance of the tracking algorithm for the Lower section (LR) was 86.756%, which was the lowest tracking performance according to the moving changes between region1 and region2 (left and right feet in the walking situation, or the crawling situation where the feet were in the back and the hands in the front).

Moreover, the algorithm performance was measured according to the child's age and sorted in the CVDS dataset. In the first group (CIn), the sample age was from four to eight months and their position was creeping, which was not one of the positions covered by this study. The body-part tracking performance in this group was 55.943%, which reflects the difficulties in labeling the head, upper and lower sections where 60.721% was achieved. On the other hand, the whole-object tracking for this group was successful, achieving 84.263% as shown in **Table 4 and Fig. 5**. The performance of this group was lower than the performance of the other groups that had walking, crawling and seated moving.

However, the WCIn group achieved the best performance in the dataset groups, reaching 91.921% in body part tracking as shown in **Table 4 and Fig. 5**. The first reason for the high performance of this group was obviously the objects, the body parts and the perfect labeling technique as it achieved 93.912% for body labeling. The second reason is that the video was captured indoors in comparison to the WCCOut group, which comprised the same objects but in an outdoor environment. The performance of this group was 87.765% for body part tracking. The performance of the body-part tracking algorithm in all the groups was 84.886% while the performance for whole-object tracking was 91.667%.

Table 5 shows the performance of the algorithm in each position for the different baby and toddler ages. Walking object tracking provided the best performance with 91.932% for body part tracking and 96.235% for whole object tracking. Crawling object tracking achieved 90.832% for body part tracking and 96.231% for whole object tracking. The last position, seated moving object tracking, achieved 89.7% for body part tracking and 93.4% for whole object tracking. **Figure 6** shows an example of the tracking technique in the three situations, i.e., the walking, crawling and seated moving positions of the child-object where both the whole object and the moving parts of the object were tracked.

3.1. Comparison the Tracking Algorithm of Child's Object Versus Tracking the Adult Object

Essentially, the tracking algorithm could be track any human object, but is its performance as same as tracking child's object? This section has the answer. The tracking algorithm is practical on INRIAPerson Dataset that has the human object in different situation (Knossow *et al*., 2008). The INRIAPerson classify to three classes as the situation of the object walking, seated moving and other moving situation.

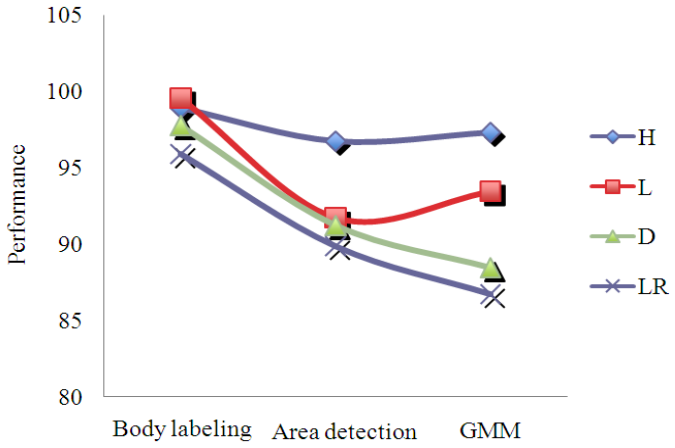

Fig. 4. The performance of the tracking algorithm in the three phases: Body labeling, body part area detection and GMM performance

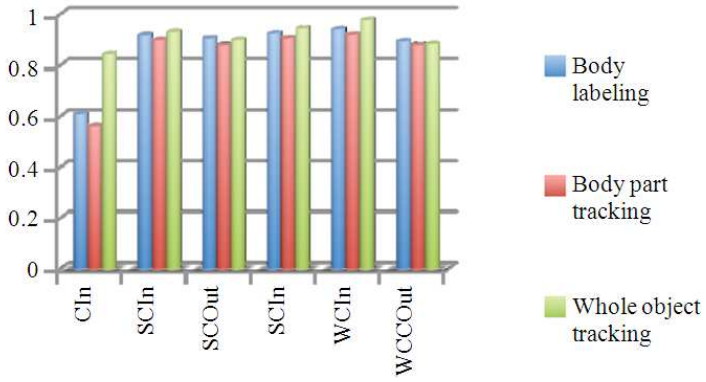

Fig. 5. Comparison of the average percentages for the body labeling technique, body part tracking and whole object tracking

Table 4. The performance of tracking technique in CVDS

	Average		
Group	Body labeling	Body part tracking	Whole object tracking
In	0.607	0.559	0.842
SCIn	0.916	0.897	0.928
SCOut	0.904	0.876	0.897
SCIn	0.923	0.905	0.942
WCIn	0.939	0.919	0.976
WCCOut	0.893	0.877	0.882

Table 5. The performance of the tracking technique for the three positions, i.e., walking, crawling and seated moving

	Average		
Situation	Body labeling	Whole object tracking	Body part tracking
Walking	0.987	0.962	0.919
Crawling	0.976	0.942	0.908
Seated moving	0.904	0.934	0.897

However, the tracking algorithm has two types of tracking: First type, labeling the body parts then tracks t it. Second type, is tracking the whole object. The tracking of the child's object success in the first type of tracking more than the second type, while the adult object success in the whole object tracking more than the body part's labeling and tracking as shown in **Fig. 7**, maybe that is because the algorithm built and designed for the child's object especially. Besides, **Fig. 7** shows the comparison of the tracking algorithm implementation in child's object and adult object, where the implementation of tracking algorithm on adult object achieved 76.667 for the tracking in whole object and 73.467 for the body parts tracking. The implementation result of tracking algorithm on the INRIAPerson dataset is lower than the implementation result of detection algorithm, on the same dataset.

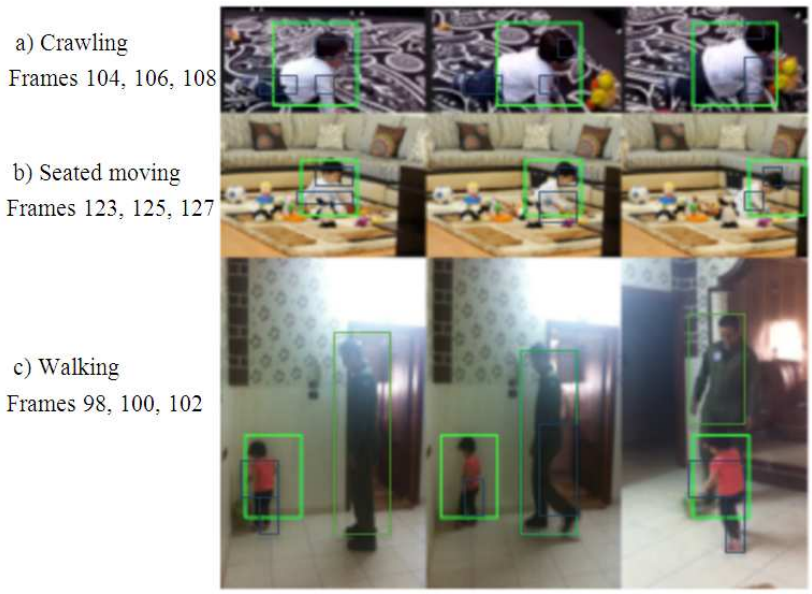

Fig. 6. Tracking a child-object in the three situations: (a) crawling, (b) seated moving and (c) walking, where the moving body parts in the three situations are detected

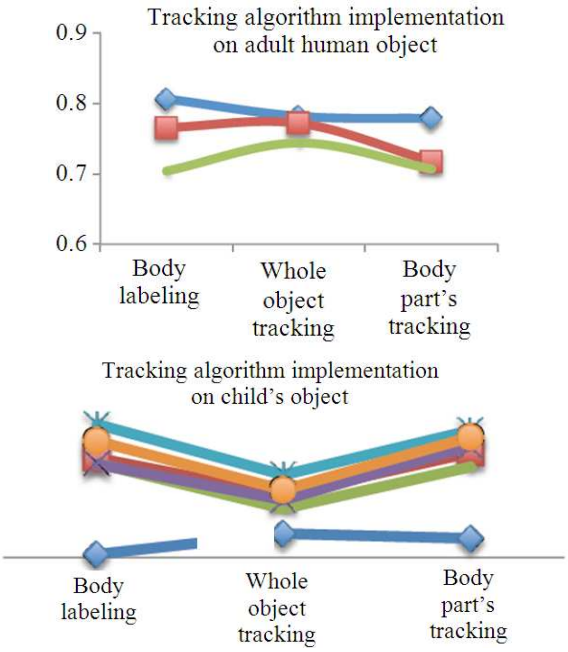

Fig. 7. Comparison of the implementation of the tracking algorithm between child's object in CVDS and adult human object in INRIAPerson dataset

Finally, the detection and tracking algorithms proved their ability in detect human object at all, but it is particularly for baby and toddler object. In this section the ability of algorithm compared with adult human object only to show that the algorithms could detect and track human object, but it is detection and tracking for the baby and toddler object is best cause this is his major.

4. CONCLUSION

In this study, we presented a robust tracking algorithm for whole child-object and his parts based on the body section labeling technique and Gaussian mixture model. The main advantage of our proposed method is that it can track three sections, i.e., the head, upper and lower sections, of baby and toddler objects both indoors and outdoors. Our experimental results show that the proposed method is robust in three situations, i.e., walking, crawling and seated moving, in different environments.

The future work: The body part labelling technics that used here are depending on (Johnson and Bobick, 2001) depth compensation method to solve the algorithm measured the object depth at an angle towards the camera. However, the algorithm could be improve by establish it is especial measuring depth technic to track the child's object in different camera depth according to the object measuring.

Completely, this study is a an approach in detecting and tracking baby and toddler object to simulate

babysitter vision, if it is findings convert to a practical software of babysitter robot vision. In other hand, constructed the hardware of babysitter robot vision as two principally moving cameras could trajectory the baby from different angles. Additionally, the progresses of this robot start from its vision to how tack care of baby how console the baby and the hard ware's of this robot are a very huge research could be benefits for researchers and parents.

5. REFERENCES

Aljuaid, H., D. Mohamad and M. Sarfraz, 2009. Arabic handwriting recognition using projection profile and genetic approach. Proceedings of the 5th International Conference on Signal-Image Technology and Internet-Based Systems, Nov. 29-Dec. 4, IEEE Xplore Press, Marrakesh, pp: 118-125. DOI: 10.1109/SITIS.2009.29

Aljuaid, H., D. Mohamad and M. Sarfraz, 2010. A tool to develop Arabic handwriting recognition system using genetic approach. J. Comput. Sci., 6: 619.624. DOI: 10.3844/jcssp.2010.619.624

Haritaoglu, I., D. Harwood and L.S. Davis, 1998. W^4: Who, When, Where, What: A real time system for detecting and tracking people. Proceedings of the 3rd Face and Gesture Recognition Conference, Apr. 14-16, Nara, Japan, pp: 222-227.

Johnson, A.Y. and A.F. Bobick, 2001. A multi-view method for gait recognition using static body parameters. Proceedings of the 3rd International Conference on Audio- and Video-Based Biometric Person Authentication, Jun. 6-8, Springer Berlin Heidelberg, Halmstad, Sweden, pp: 301-311. DOI: 10.1007/3-540-45344-X_44

Kanhere, N.K., S.J. Pundlik and S.T. Birchfield, 2005. Vehicle segmentation and tracking from a low-angle off-axis camera. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 20-25, IEEE Xplore Press, pp: 1152-1157. DOI: 10.1109/CVPR.2005.365

Knossow, D., R. Ronfard and R. Horaud, 2008. Human motion tracking with a kinematic parameterization of extremal contours. Int. J. Comput. Vis., 3: 247-269. DOI: 10.1007/s11263-007-0116-2

Rosales, R. and S. Sclaroff, 1999. 3D trajectory recovery for tracking multiple objects and trajectory guided recognition of actions. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 23-25, IEEE Xplore Press, Fort Collins, CO. DOI: 10.1109/CVPR.1999.784618

Sarfraz, M., H. Aljuaid and D. Mohamad, 2011. Evaluation approach of Arabic character recognition. IJCVIP 1: 58-77. DOI: 10.4018/ijcvip.2011040105

Sasaki, H., T. Fukuda, M. Satomi and N. Kubota, 2009. Growing neural gas for intelligent robot vision with range imaging camera. Proceedings of the International Conference on Mechatronics and Automation, Aug. 9-12, IEEE Xplore Press, Changchun, China, pp: 3269-3274. DOI: 10.1109/ICMA.2009.5246241

Shi, J. and J. Malik, 2000. Normalized cuts and image segmentation. IEEE Trans. Patt. Anal. Mach. Intell. 22: 888-905. DOI: 10.1109/34.868688

Stenger, B., V. Ramesh, N. Paragios, F. Coetzee and J.M. Buhmann, 2001. Topology free hidden markov models: Application to background modeling. Proceedings of the 8th IEEE International Conference on Computer Vision, Jul. 7-14, IEEE Xplore Press, Vancouver, BC., pp: 294-301. DOI: 10.1109/ICCV.2001.937532

Torralba, A., K.P. Murphy and W.T. Freeman, 2007. Sharing visual features for multiclass and multiview object detection. IEEE Trans. Patt. Anal. Mach. Intell., 29: 854-869. DOI: 10.1109/TPAMI.2007.1055

Vedaldi, A. and S. Soatto, 2006. Local features, all grown up. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 17-22, IEEE Xplore Press, pp: 1753-1760. DOI: 10.1109/CVPR.2006.176