www.icgst.com

Shopping behaviour analysis in videos

R. Sicre, H. Nicolas
LaBRI, University of Bordeaux
351 Cours de la libération, 33405 Talence Cedex, France
MIRANE SAS
16 rue du 8 mai 1945, 33150 Cenon, France
[sicre, nicolas]@labri.fr

Abstract
In this paper, we propose a method that analyzes human behaviour in a shopping context. This method is based on motion detection and object tracking. Our research is developed in order to build a real-time application that will analyze people's behaviour while shopping. Specifically, we detect customer interactions with specific products.

Keywords: *computer vision, real-time object tracking, event detection, video surveillance, marketing.*

1. Introduction
Computer vision is used in various applications including: mobile robot navigation, active surveillance or even video games.
Many applications use object tracking. The performances of these techniques, based on regions or objects, allow us to analyze the semantic context of a scene that is the main topic of the paper. The applicative context is the use of video surveillance cameras for marketing purposes. Various software systems are available today allowing to track customer in order to cumulate statistical information regarding their habits. At the same time, Digital Media are more and more used to display advertisement in shopping malls. Several software systems measure the audience in front of the screen based on face detection. However, these systems are not really interactive and just display advertising clips one after another.
The study introduced in this paper is in the same context and aims at adapting the clips to the behaviour of the present customer. It requires a scene analysis step that leads to an interpretation of the customer behaviour. Typically, we detect products grabbed in a known area in real-time. The detection of this kind of event result, for example, in playing a clip related to the specific product.
After a short review, the paper describes our system: first the method used to detect and track objects, then the analyze that detect interaction with products. We discuss the position of the camera before showing the results. Finally the conclusion presents future work.

2. State of the art
This section presents different techniques for motion detection, object tracking and behaviour analysis. We focus on fast methods in these three categories of previous work, because the final application has to work in real-time. Also, the system is used indoors, so we will not face major light changes. We, therefore, focus on methods designed to work under these conditions. For an overview of the field, the reader can refer to [23], [6] and [12].

2.1. Motion detection
Motion detection aims at distinguishing the background from the moving objects. The techniques can be separated in two categories depending if there are using a background model or not. Without model, the algorithms are based on temporal differencing, that make the difference between two or three consecutive frames [19], or calculation of optical flow [1]. These methods can be really fast, but the detections are very sensible to noise.
When a model is available, we can classify method depending on the type of model used. The model can be pixel based, local or global.

2.1.1. Pixel based model
These methods associate to each pixel of an image a value or an intensity function that gives the appearance of the background. We only use the measurement taken on the specific pixel. We present here three main methods:
Background subtraction is a simple method that uses an image of the background without objects as a reference [22].
Gaussian model is a popular method that consists in using one or more Gaussian distribution [17].
Segmentation methods, like the « Mean-shift » algorithm, is a newer technique used in pattern recognition. It classifies complex distribution of data into clusters [5]. However it is computationally expensive.

2.1.2. Local model

The previous methods only use pixels self measurements and are not very precise due to the nature of the motion and the relative texture in the scene. To face these problems, it is possible to use the neighbourhood of a pixel and calculate a similarity measurement. Specifically, these methods calculate if a bloc of pixels belong or not to the background. Each bloc needs to have a descriptor that is based on its gradients or a colour code. These methods can also differentiate moving objects from their shade and the background [15].
As this detection method is computationally expensive, it is usually combine with a pixel based method to reduce the area where the local detection is applied.

2.1.3. Global model

Global methods use the entire image at each moment and build a model of the entire background.
The k-means algorithm build k background model. A pixel based detection is then applied with the different background to select the best model [20]. Hidden Markov Model (HMM) can also be applied to detect global changes in the scene. These changes are usually related to outdoor scene [16]. The «Eigen background » method is a different use a different way to represent the image. It creates a set of appearance model that only possess a few significant dimensions. This method allows an efficient global detection [13].

The motion detection phase is really important. The following processes, especially the tracking part, are directly depending on the good quality of the detection results. In our study we chose a pixel based model that offers a good compromise between quality and speed.

2.2. Object tracking

After moving regions are detected, regions are tracked over the frame sequence. Some detection and tracking methods use the same tools, therefore methods can merge. First category of methods builds a descriptor for each region and matches them with regions descriptors of another frame. Also, Regions can be directly compared to models created for different type of regions. As in [6], we sort methods into four categories. Tracking can be based on region, contours, features, and model.

2.2.1. Region based tracking

This method is intuitive and aims at identifying connected regions corresponding to each object in the scene. Regions are usually the result of the motion detection. The problems in this method are that several regions can correspond to one observed object. Regions are usually described based on their colour and gradient [10].

2.2.2. Active contour based tracking

This process uses the contour of each region. The shape of the detected regions is matched from one frame to another. People can be tracked using deformable contour models. In particular, human are considered as articulated objects so that important deformation occur depending on the camera position [2].

2.2.3. Feature based tracking

Feature based tracking do not aim at tracking an object, a person, as one entity. It uses distinctive features that can be local such as points, line, curves, and corners [4] or global like perimeter, surface, colour, and position [14]. These data are used to describe and then match the objects.

2.2.4. Model based tracking

This last method allows various ways of modelizing objects: articulated skeleton [8], 2-D contours [7], 3-D volumes, etc. For a new image, detected regions are compared to different models previously built. Models are updated with new information.

The tracking method developed in this paper is based on regions features and regions its elves.

2.3. Behaviour Analysis

This section comes after object tracking. We are especially interested in observing humans that are deformable objects. The goal here is to analyze and recognize motion samples in order to draw high level conclusion. This analyze is made in two steps, description and recognition of specific actions. In fact, before recognizing any action, the fundamental problem is to describe each possible action, find measures and the way there varies along time.
Once the training is realized, there are several methods that can link the data. The Dynamic time warping is a simple process with robust performances [18].
HMM can also be applied [3]. They are more efficient while analyzing data as a sequence, but needs a learning phase before the classification.
Neural network can deal with important amount of data varying over time. There are used in human posture recognition [21] and lips reading [11].

3. System architecture

Figure 1 shows the different module of the application. Motion detection is followed by tests on the areas. Then regions are merged and tracked. Finally the tracked objects are analysed in order to interpret their behaviours. The major phases are detailed in the coming paragraphs.

3.1. Detecting moving objects

The background model is obtained by calculating the mean of several images while no object is in the scene.
The images (current and model) are converted into YUV colour space. The images are compared to detect foreground objects on each channel. Chrominance channels receive a higher weight than the luminance channel in order to reduce the effect of shadows. However, without any explicit model of the illumination variations, small illumination variations on highly textured areas can lead to false positives.
We apply morphological filters on the image of detected motion. Closing improves the regions shape and fills the holes. Opening reduces the detection noises. An example of detection is shown, Figure 2. We can see classical challenges as textural similarity between moving object and background. We need to merge some regions.

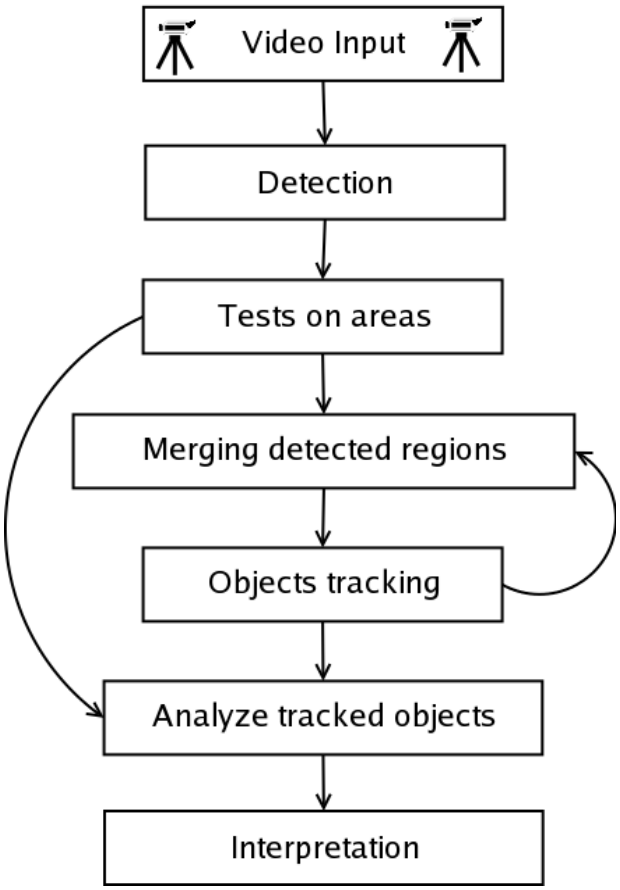


Figure 1. Functional diagram of the system

3.2. Regions merging

As several detected regions can correspond to the same person, we developed a merging module based on various criteria that correspond to distance and relative position of regions. For each detected region, a bounding box is calculated and merging is based on the three following criteria.

- Covered bounding boxes: when two boxes overlap, the regions are merged
- Close by regions: when the boxes are in the same neighbourhood, regions are considered as "potentially merged".
- Regions located in the same vertical axis: As we are essentially detecting people, when a person is divided into several blobs, regions often share the same vertical axis. When regions are detected in such a configuration, we can presume that they correspond to the same person. Specifically, if the centre of gravity of a region is situated between the two extrema (left and right) of another one, these two regions are considered as "potentially merged", see Figure 2 (e).

When merging is potential or uncertain, the tracking phase validates the merge or not.

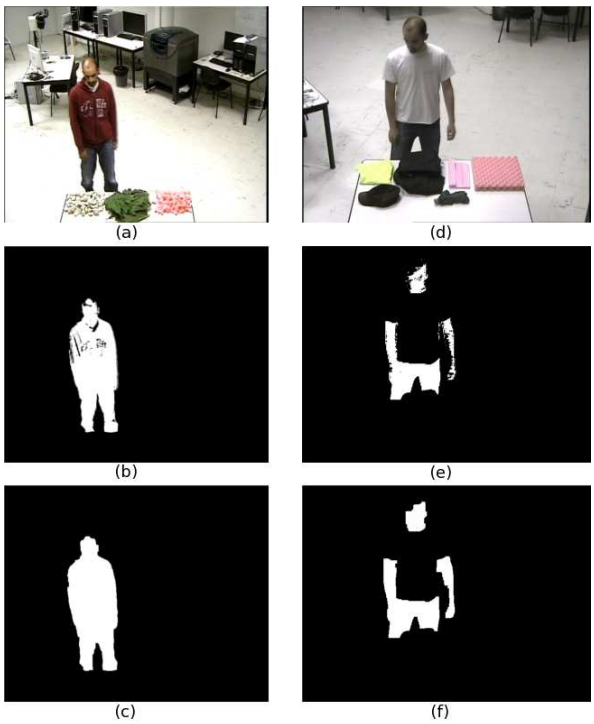


Figure 2. Input images are (a) and (d). (b) and (e) are motion detection. (c), (f) are the final results after morphological filtering.

3.3. Object tracking

After detecting regions, the system calculates a descriptor for each of them. The descriptor is based on its position, size, surface, and first and second colour moments. The first moment corresponds to the mean of the colour (respectively on each channel: RGB) and the second moment is the variance. These descriptors are matched from one frame to the next in order to track regions.

We use a similarity measurement [9] that works as follows: for each measurement i of the descriptor M_A^i of the first image region *A*, we have *k* regions $B_0,\ldots,B_k$ of the second image, with (for all of them) a i^{th} measurement $M_{B0}^i,\ldots,M_{Bk}^i$. All these measurement are compared (by calculating Euclidean distance) to M_A^i in order to find the closest one. The closest measurement receives a vote that suggests the link between the region *A* and the corresponding region of the second image. All these votes are collected to determine the closest region B_j to *A*.

Once this process is executed on each region of the first image, the same technique is applied in the other direction. In other words, for each region of the second image, the algorithm looks for the closest region in the first image.

In order to match two regions: A_r and B_l coming from two different images, we find if B_l is the closest region of A_r and if A_r is the closest region of B_l.

It is interesting to note that if there is only one or two regions, regions with very different descriptor can be

matched. On the other hand if there are many regions, the criteria are harder to satisfy.

3.4. behaviour interpretation

In our study, the background is divided in two principal areas: interest areas, where various products heaps are located, and the rest of the scene. These areas are obtained after a manual initialization phase. Since the system knows the interest areas, regions are classified in two categories whether their gravity centre belongs to an interest area or not.

Region previously tested that do belong to an interest area are considered as product that were removed from the products heap, probably taken by a customer. This information concerning the scene is relevant. However this information is known once the action is done. Customer behaviour is usually to think, look at the price, and then take a product. We want to know what they are interested in as soon as possible.

For this reason, we detect customer occluding interest areas. Typically regions are detected and tracked in the scene. If a region does not belong to the interest areas, this region corresponds to a client shopping. At some point, a part of the client can be detected as covering an interest area. When a surface that is bigger than a hand theoretical size is detected, then a new event concerning the customer behaviour is detected: the customer is about to grab a product. As we know where the camera is located, we set a hand size to a certain amount of pixels in the image.

Other clues concerning the interest of a customer can be deduced from the customer position. In fact, a customer is likely to pick up products that are close to him. The direction the customer is heading to can be of interest too. For our application, information regarding a customer location is less significant than the detected behaviour, we just present.

4. Camera position

The camera position is an important factor. The case we study is simple: taking object on a table. Thus, our choice of camera position is quite flexible. Knowing that the application will be used for different kind of racks, we want to optimise this position.

The basic question to solve is: “What do we observe?” “The clients or the products?”. Although our most relevant observations come from the products, the clients offer much more information. Various trials have been made, changing the camera position. We noticed that when people or products cover a significant amount of pixels, the results improve due to the resolution increase. However, the merging process has to merge region that are far one from another and can show some limitations when several customer are in the scene. It is also important to see the entire customer while he is picking up products.

We also noticed that when the person takes more space in the image, the field of view is then reduced and the noise generated by the camera becomes significant. Focus automatically changed generates blurs. The auto-white-balance causes problems too. In fact, when an object appears that significantly changes the contrast, pixels values of the entire image are significantly modified. It is important to fix these camera functions or to effectively compensate for them. Some cameras can possess an auto-iris that allows letting more light into the sensor. It needs to be controlled or compensated.

Knowing these side-effects, we fixed the auto-white-balance as well as the iris of the camera.

5. Results and limitations

Our system has been tested on various sequences, especially some taken in the laboratory facilities.

5.1. Object tracking

The motion detection module gives good visual results, see Figure 2 that help the tracking phase.

The Figure 3 (a), 3 (b), and Figure 6 (a) allows us to visualize tracked people trajectories, during the videos sequences shown in Figure 4 and 5. The respective errors compared to ground truth is shown Figure 3 (c), 3 (d) and 6 (b). Ground truth was obtained by hand by getting at each frame, the centre of gravity position (human hips level). It is interesting to note that for Figure 3 (e) starting at frame 100, the error increases only along the Y axis. The same phenomenon appears at frame 150 for Figure 3 (f). We note that it corresponds to the exact moment when people get close enough to the table to have their legs occluded. Entering and exiting the scene gives noise as long as the person is not entirely in the scene.

There are several limitations to the system. Most of them come from the merging process. It gives good results in simple cases. However, if two people are in the scene and very close one to another, they might be considered as one single object, especially when they are static. In these cases, separating the two people that are similar and tracking them is challenging. We note that we reach the limitations of the merging process faster than the tracking limitations. Figure 3 (f) shows a peak of the error along X and Y between frame 390 and 410. The person takes an object on its side on the floor and the tracking loses track of a part of the person over a few frames, see Figure 4 frame 403.

The tracking works correctly in a case of a single object divided in several regions. Such a division is due to people clothes or partial occlusions of the scene, see Figure 4 frame 25. Tracking is correct for two or three people in the scene as long as they are not close one to another, see Figure 5. However, the system does not handle occlusion between people.

5.2. Picking up products

The result for detecting taken objects, or a region in interest area, is not precise. In fact, it is difficult to detect major changes when a product is taken from a heap of identical products. Also false positives can occur if the products are highly textured, due to camera adjustments. Finally, false negatives occur when the person occludes the taken object. However it does not affect the results, because the detected product merges with the person and then increases the occluding surface on the interest area.

Products about to be taken offer much better results. However false positives can occur due to shadows or

when a person covers several interest areas at the same time. The most occluded area is selected in such a case.
When these two measurements are combined, they offer a good interpretation of a picking up behaviour in a specific area. Figure 3 shows three different product areas, numbered from 1 to 3 from the left to the right, the corresponding video sequence is shown Figure 4. Figure 6 (e) and (f) give correct results in a much complex scene with multiple and simultaneous interactions.

5.3. Execution times

The final application of this project need to work in real-time, we use simple and fast methods. The program is tested on a computer with two Pentium IV 3Ghz with 1 GB of RAM. The system process 6 to 10 frames per second depending on the video resolution, see Table 1.

Execution times \ Resolution	640*480	704*576
Motion detection	0.07 s	0.1 s
Saving regions	0.025 s	0.035 s
Tests and merging regions	0.005 s	0.01 s
Objects tracking	< 0.005 s	< 0.005 s

Table 1. System execution times for videos of different resolution.

6. Conclusion

This paper presented a simple method to track objects in real time. It also analyzes the behaviour of tracked objects such as picking up object in known areas.
In order to improve the system, we first want to improve the high level analysis. We only use a few measurements to give clues concerning behaviours. We want to group these data in order to detect full scenarios
We can also improve the detection method with a better background model and the tracking method that manages occlusion and filters data.

7. References

[1] J. Barron, D. Fleet and S. Beauchemin, "Performance of optical flow techniques," *Int. J. Computer Vision*, pp. 42–77, 1994.

[2] A. Baumberg and D. C. Hogg, "Learning deformable models for tracking the human body," *Motion-Based Recognition*, Kluwer, pp. 39–60, 1996.

[3] M. Brand, N. Oliver and A. Pentland, "Coupled hidden Markov models for complex action recognition," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 994–999, 1997.

[4] B. Coifman, D. Beymer, P. McLauchlan and J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," *Transportation Res*. pp. 271–288, 1998.

[5] B. Han, D. Comanicu, L. Davis "Sequential kernel density approximation through mode propagation: Applications to background modelling," *6th Asian Conf. on Computer Vision*, 2004.

[6] W. Hu, T. Tan, L. Wang, S. Maybank "A survey on visual surveillance of object motion and behaviors," *IEEE Transaction on systems, man, and Cybernetics*, pp 334 – 352, 2004.

[7] S. Ju, M. Black, and Y. Yaccob, "Cardboard people: a parameterized model of articulated image motion," *IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 38–44, 1996.

[8] I.A. Karaulova, P.M. Hall, and A.D. Marshall, "A hierarchical model of dynamics for tracking people with a single video camera," *British Machine Vision Conf.*, pp. 262–352, 2000.

[9] J. Matas, O. Chum, M. Urban, T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," BMVC, pp 384-393, 2002.

[10] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Computer Vision Image Understanding*, pp. 42–56, 2000.

[11] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel, "Toward unrestricted lip reading," *Int. J. Pattern Recognit. Artificial Intell.* pp.571–585, 2000.

[12] T. Moeslund, A. Hilton, V. Kruger, "A survey of advances in vision-based human motion capture and analysis", *Computer vision and image understanding*, pp 90-126, 2006.

[13] N. Olivier, B. Rosario, A. Pentland, "A Bayesian computer vision system for modelling human interactions," *IEEE Trans on Pattern analysis and machine intelligence*, pp 831-843, 2000.

[14] R. Polana and R. Nelson, "Low level recognition of human motion," *IEEE Workshop Motion of Non-Rigid and Articulated Objects*, pp. 77–82, 1994.

[15] J. Rittscher, J. Kato, S. Joga, A. Blake, "A probalistic background model for tracking," *6th European Conf. on Computer Vision*, pp 336-350, 2000.

[16] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, J.M. Buhmann, "Topology free hidden markov models: application to background modelling," *IEEE, conf. on Computer Vision*, pp 294-301, 2001.

[17] C. Strauffer, W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp 2246-2252, 1999.

[18] K. Takahashi, S. Seki, H. Kojima, and R. Oka, "Recognition of dexterous manipulations from time varying images," *IEEE Workshop Motion of Non-Rigid and Articulated Objects*, pp. 23–28, 1994.

[19] Y.L. Tian and A. Hampapur "Robust salient motion detection with complex background for real-time video surveillance," *IEEE Workshop on Motion and Video Computing*, pp 30-35, 2005.

[20] K. Toyama, J. Krumm, B. Brummit, B. Meyers, “Wallflower: principles and practice of background maintenance,” *IEEE Conf. on Computer Vision*, pp 255- 261, 1999.

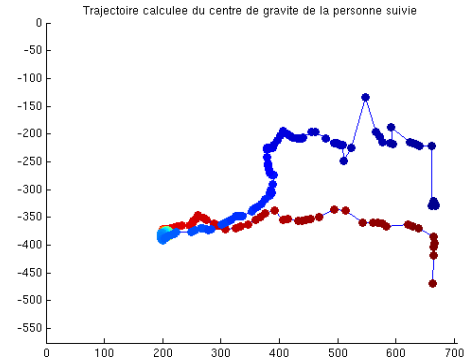
[21] M. Yang and N. Ahuja, “Extraction and classification of visual motion pattern recognition,” *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 892–897, 1998.

[22] T. Yang, S.Z. Li, Q. Pan, J. Li, “Real-time and accurate segmentation of moving objects in dynamic scene,” *ACM Int. Workshop on video Surveillance & Sensor Networks*, pp 136-143, 2004.

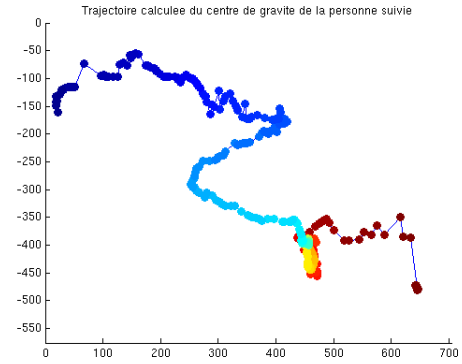
[23] A. Yilmaz, O. Javed, M. Shah, “Object tracking: a survey”, *ACM Computing Surveys*, 2006.

R. Sicre received a Master degree in Intelligent Systems from the University of Toulouse III, France, in 2007. During his master he studied at the University of Plymouth, England, at the University of Calgary, Canada, and at the Technical University of Berlin, Germany. Since 2008, he is working toward the Ph.D. degree in the LaBRI at the University of Bordeaux 1 in collaboration with MIRANE S.A.S.

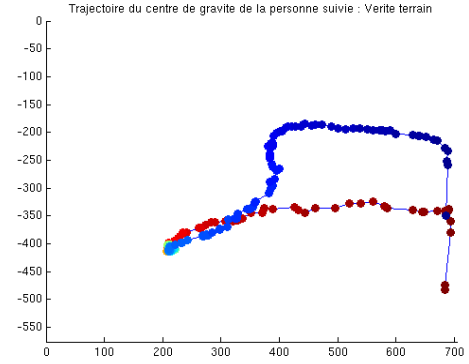
H. Nicolas graduated from the INSA of Rennes in 1988, and received his Ph.D. degree in computer science from the University of Rennes in 1992. For his PhD, he worked at IRISA/INRIA of Rennes, France. The subject of his thesis was related to video compression. From 1993 to 1995, he worked at the 'Swiss Federal Institute of Technology of Lausanne' (EPFL, Switzerland) as the head of the research group on Digital Television in the Signal Processing Laboratory. From 1996 to 2005, he worked as a senior researcher at INRIA of Rennes. Since September 2005, he is a full professor at the University of Bordeaux 1, France.



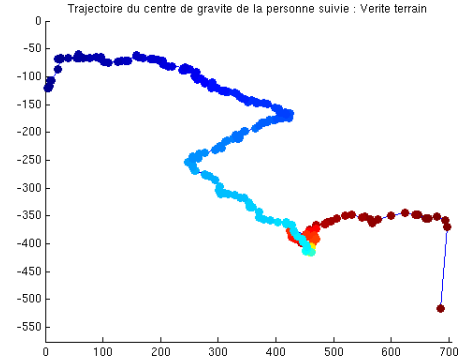
(a)



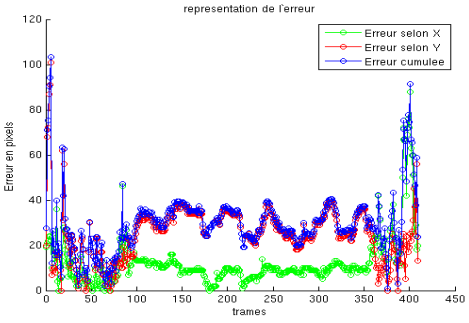
(b)



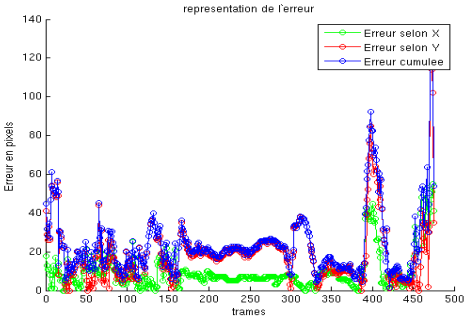
(c)



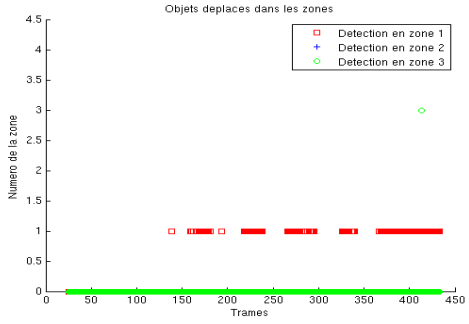
(d)



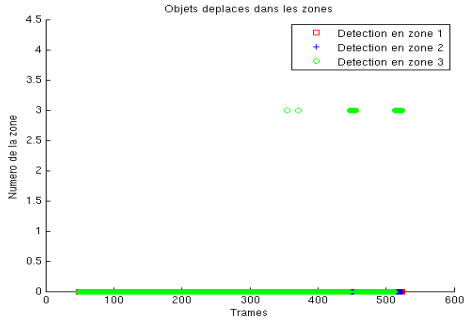
(e)



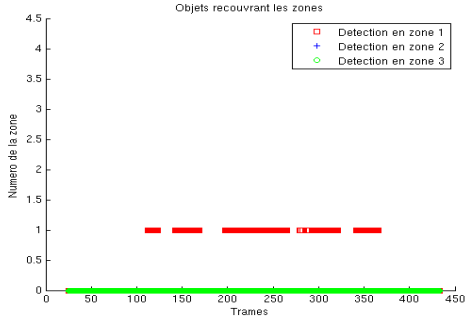
(f)



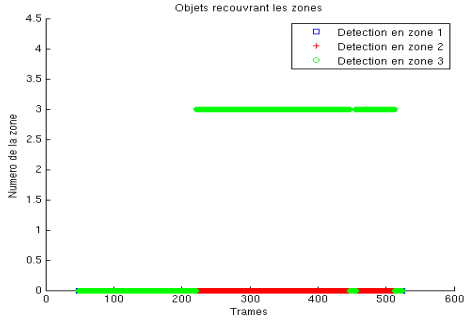
(g)



(h)



(i)



(j)

Figure 3. Tracking results. (a) and (b) are the person gravity centre trajectories. (c) and (d) correspond to ground truth. (e) and (f) are the error in pixel along X axis in green, along Y axis in red, and cumulated in blue. (g) and (h) represent motion detected in an interest area. Finally, (i) and (j) correspond to region covering an interest area.

Figure 4. The first line correspond to the video that give the results of the first column in Figure 3, the second video sequence give the results of the second column.

Figure 5. Video sequence where two people are successfully tracked, images correspond respectively to frame: 14, 19, 22, 36, 41.

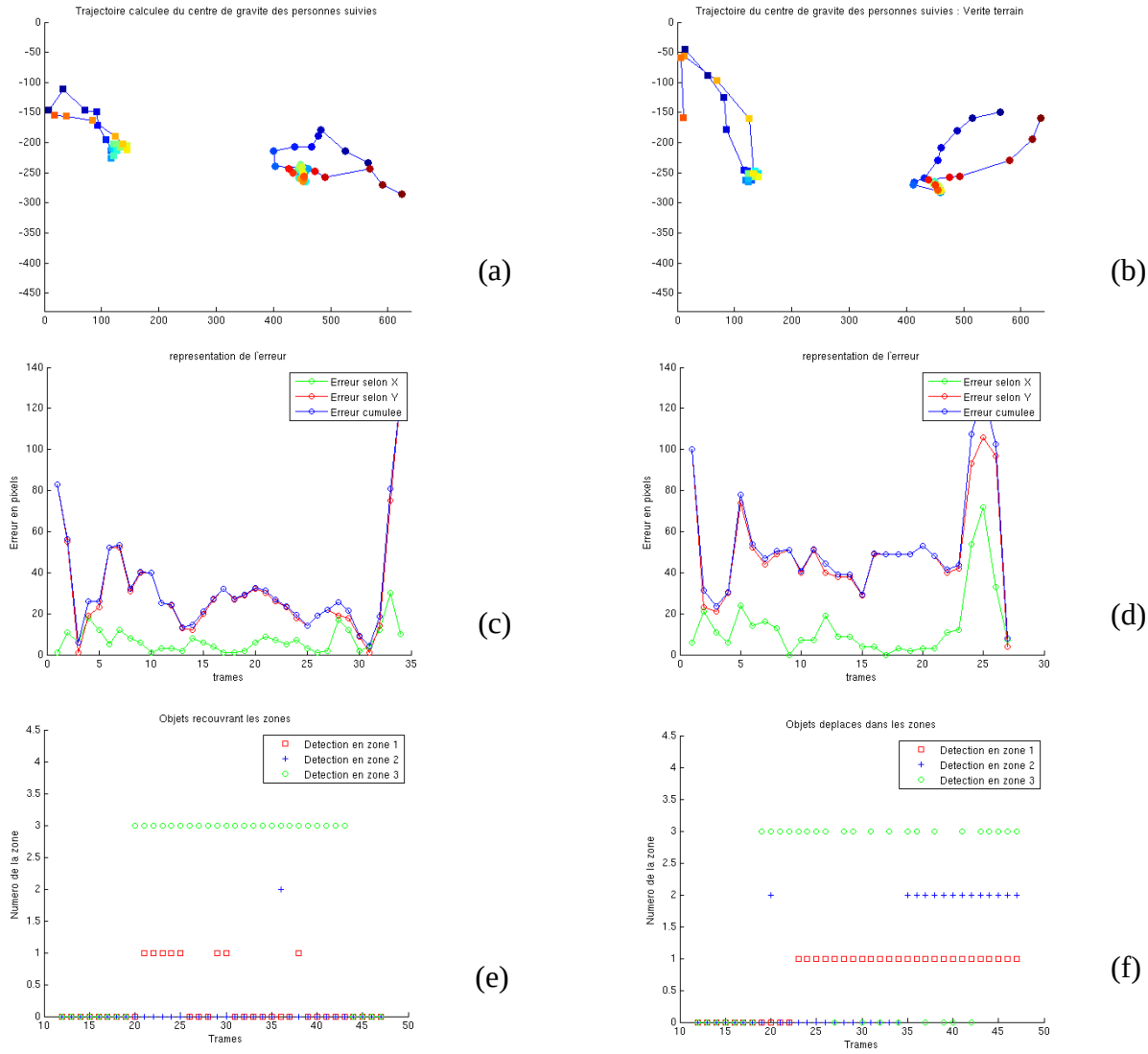


Figure 6. (a) and (b) are the trajectories of the two people: calculated and ground truth. (c) is the error for the person on the right and (d) for the person on the left. (e) represents people covering products and (f) taken objects.