Consumer buying metrics extraction using Computer Vision techniques

J. P. D'Amato[a,c,1], C. GarciaBauza[a,c] , E. Rinaldi[b]
[a]*InstitutoPladema, Fac. Cs. Exactas, UNCPBA., Tandil, Argentina*
[b]*Fac. Cs. Economicas, UNCPBA. Bs.As.,Tandil,Argentina*
[c]*CONICET, Bs. As. Argentina*

Abstract. Information about how people move inside a market and interact with products helps to take important decisions in marketing strategies and products layouts. Using digital video cameras, people inside a location could be detected and counted. Once a person has been detected, it is possible to study whether he interacts or not with a product, evaluating the interaction points, the purchase average time, the number of acquired products; among others. All this data that could only been detected by observation methods, lets companies to make considerations about how convenient is selling certain products. In this work, we present a system that applying computer vision techniques can analyze and generate reports in real time. These reports help decision-making in pricing policies, customers' preferences and product placement. A scalable architecture was developed to manage multiple capture points and store in a database. To validate the classifications results, we filmed over 100 hours of video in a supermarket and several metrics were extracted that corresponds with actual values provided by a sale system.

Keywords. Computer vision, behavior detection, shopping metrics.

1. Introduction

Video systems are used to monitor people activities in many situations [1] and their demand for these systems has rapidly increased in the last years . They are used mostly in security, but are also been an important tool in sports (e.g. the well-known Hawkeye), in entertainment (e.g. in augmented reality) and even in shopping activities. In the last case, they provide important information, like the amount of people moving and carrying object through a location. Robust motion detection algorithms to track people and analysis methods to recognize different actions are required.

At the same time, marketing has suffered a radical change in the way sellers try to attract and provide services to customers. An overall accepted strategy to improve sales is to exhibit products in the shelves with a particular arrangement using Planograms[2]. Products are strategically accommodated according to price, shop priority or genre. Planograms serve to this purpose, but the customers' interaction with a product cannot be measured. This interaction, also called as "pick-up", is an important indicator for a brand and can only be extracted by observation. Total purchases can be directly extracted from a commercial system, for this reason they are not an item of interest in our research.

[1]Corresponding Author.

To observe and verify the effectiveness of the deployment, it can be used personal satisfaction inquiries, but they need a great effort in designing and not always people accept them. On the other side, video analysis techniques are an interesting approach because they are not invasive and do not influence in the customers decisions. There are some commercial tools based on video analysis that provides such information as presented in [20].

At the same time, the response speed and confidence in the information extracted are important to achieve a competitive difference. The captured information should be available and accessible by different users (repositories, managers, business analysts) at a relatively low cost. Even more, the system that delivers this information should also adapt to the different features of the studied place like size, amount of shelves, and distance between shelves, among other. As the field of view of the camera is limited, it is important to have multiple simultaneously points if we want to evaluate a great area.

The purpose of this work is to design and implement analysis techniques that automatically identify and record customers and their interaction with products. We develop an image analysis system that can detect people and their movements while their identities are preserved. It is proposed to mount cameras overhead and extract aerial images, using either RGB or depth sensors. This layout does not affect the daily commercial activity but introduces a challenge of processing the images in a different way. In **Fig. 1** is shown an example of two different camera layouts.

To make the system scalable, processing and decentralized storage architecture is presented. Each capture device is connected to a processing unit (a computer), which is accessible via a server connection using a reverse proxy scheme [17]. A central server summarizes the results obtained using different displays and charts, while multiple video sources in different positions are managed.

Figure 1 – (left) (right)Vertical view of the shelves

This work is organized as follows: in Section II some background and required features are explained. Section III and IV describe some implementation details. In Section V, it is shown a study case and results. Finally, the conclusions can be found in Section VI.

2. BACKGROUND

Consumer buying behavior represents "the decision process and acts that make people buy and use products" [3]. Monitoring this behavior is of particular interest for both the

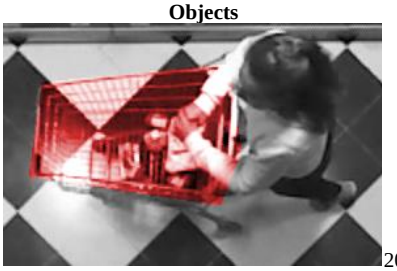
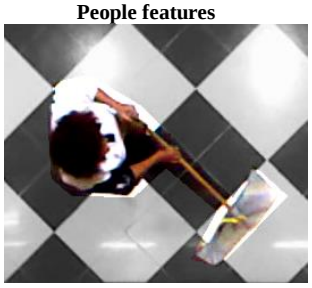
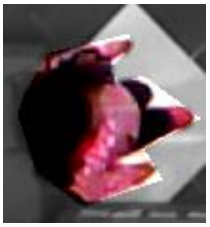
academia and the private sector. There are works that cover this topic, mainly applying image analysis techniques. The work of [4] deals with the analysis of people in general, with major applications in security and surveillance, but also in the consumer market. Other attempts to analyze purchases are based on *in-situ* observation [5].Both proposals are too general and deals only with motion detection.

In [20], consumers behavior is studied to find motivations and conscious and unconscious influences, using video records and interviews with consumers but it is not automatized. A study more alike ours is proposed in [16]. In this work, the authors propose integrating different media data and analysis, but they only present the main idea and do not show any real cases.

2.1. People counting and environment issues

As it was mentioned above, people tracking is a key component in these studies and several works deals with it. However, when a system is deployed in public open places, some particular issues overcome that presents additional challenges [19]. These challenges arise from a variety of situations when data is captured, such as uncontrolled lighting or noise changes, or from people having unpredictable actions. Moreover, it is not always easy to discriminate how many people are in front of the camera. In [15], they adopt a vertical camera and propose a counting people algorithm but they do not consider people carrying objects or holding kids. Some cases are shown in **Table 1**. We use their idea to identify people and extend it to consider more cases and we also deal with interaction analysis.

Table 1. Some commonly situations that should be treated

Objects	People features	Kids
 20 13-10-23 : 20:31		
		

To discriminate people and reduce capture noise, it is necessary to apply special processing techniques images. Next, we will enumerate the best known techniques, some of which are used in our proposal:

2.1.1. Image pre- processing

In general, images that are captured in a controlled environment are free of lighting fluctuations, noise or other disturbances. In real situations, it is necessary to improve the quality of the images applying per-pixel algorithms before more complex analysis is carried out [21]. Among these techniques, it can be found smoothing techniques, erosion / dilation, threshold, binary masks, among others.

2.1.2. Motion Detection

Behaviors are associated with some kind of movement. For this reason, detect movement in a scene is the first and most important step toward understanding behavior. Motion detection is designed to detect non-static parts of a scene by comparing two or more consecutive frames. The techniques used for motion detection can be divided into several categories: background subtraction [7], temporal differentiation [8], or optical flow estimation [9]. Most of the following analyses are highly dependent on this stage. In [18], a work from this group implementing a parallel version of a subtracting background was presented.

2.1.3. Modeling background and objects

Background modeling is very important for the movement detection, providing a description of the scene that can aid in the interpretation of the observed data. In the studied cases, the section store can be divided into two zones, such as product and walking areas. The original image could be partitioned in areas called regions of interest (ROIs). ROIs are manually selected. These regions could be rectangular areas or binary masks. Modeling background correctly can greatly reduce the cost of computing and help to eliminate false positives.

Even more, certain objects could be identified and removed from the scenes. In the work [15] a particular object segmentation algorithm is used, applying a depth filter.

3. EVALUATION METHODOLOGY

To design and implement the automatic study of people shopping behavior, an analysis methodology is carried out in the following stages:

• *Identification* and monitoring (or tracking): all the people who pass through the analysis area should be detected. Objects that are not of interest should be discarded.

• *Interaction*. This stage is considered when the person interacts with a product. Here two interaction behaviors are possible: the person inspects the product and returns it to its place or buys it.

• *Classification* of behavior. We consider recording the activity of people that spend a certain time in the monitoring area.

Each of these stages is associated with a different processing strategy that we show in the following sections.

To ease the analysis, background modeling should be performed. As in the propose layout, the camera is mounted on top of the shelves, and four sections could be identified, as shown in **Fig. 2**.Input/output sections are used to start or stop a tracking. Shelves section is used to determine where an interaction was carried out. Transit section is always monitored.

Figure 2 – ROIs set for a typical layout

3.1. People detection algorithm and tracking

Vertical capture setting reduces the feeling of being "observed" as people do not realize that they are been analyzed. At the same time, identity is preserved because face is not captured as it is of no interest for our studies.

To carrying out people detection, it was first evaluated whether to use images in RGB colors or depth images. A device like the Kinect® provides both kinds of data. As it was discussed before, depth images help to quickly detect moving objects and give precise distance information. Moving detection filters based on RGB camera are very sensitive to changes in local illumination, and are less effective when people appears together. Hence, to detect people we used depth images.

Before detecting people, a threshold distance filter is applied to remove far points that generally correspond to the floor. At the same time, it could be used to discard all objects lower than 90cms height, like A and B.

Then, to discriminate between one or many individuals, a variant of the algorithm called "water fill" [15] was implemented. Here, we propose to seek "higher pixels" that corresponds to the head and then start detecting other parts like shoulder and arms, searching from top to bottom. When several people are presented in the image, pixels corresponding to other parts of the body could merge, but the head could be always isolated. Kids been hang up could be also detected and counted with this algorithm. They will be discarded later.

This classification is presented in Fig.3.

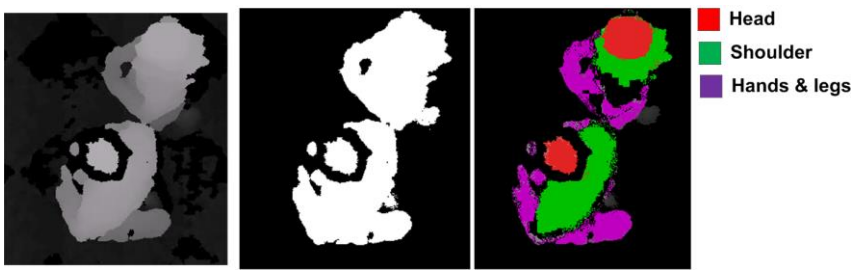


Figure 3 – People detection (left) depth image (center) removing floor pixels (right) body parts estimation

For start tracking people, the I/O areas are first analyzed, as indicated in **Fig.2**. If a new object is detected in this part of the scene, it is assigned an ID with the frame number. In the following frames, it is observed if other objects were detected and tagged with the same ID if they are closed enough. If an object appears in several successive frames, it is classified as a *person* and a new record is generated, otherwise it is discarded. A person is no longer tracked when he left the image. For this purpose, we estimate the position in the next frame using the trajectory. If we find he is outside the frame, he is discarded for the next analysis. Finally, with depth images, some extra information is stored like people height or shoulder width that can be used later. In some cases the tracking algorithm fails, for example when a person crouch is lost because of the height filter or when they leave the analysis area, and then immediately return back.

3.2. .Interaction detection algorithm

The interaction begins when a person touches the product. This interaction is generally easy to detect using regions of interest near the shelves, in which a background subtraction algorithm [18] is applied. This type of processing is effective but produces many false positives (for example, if a person goes walking through the detection zone).To reduce these errors, a combined analysis is taken.

The analysis starts when a motion is detected in front of the shelves. If the movement is repeated in successive images, it possibly corresponds to an interaction and it is stored. Then, it is evaluated if the interaction belongs to an identified person checking if the pixels are connected. At the same time, the interaction position is estimated using the distance information and position inside the area, making a correspondence with a front view of the shelve previously calibrated.

Figure4 describes the design of the system used here presented.

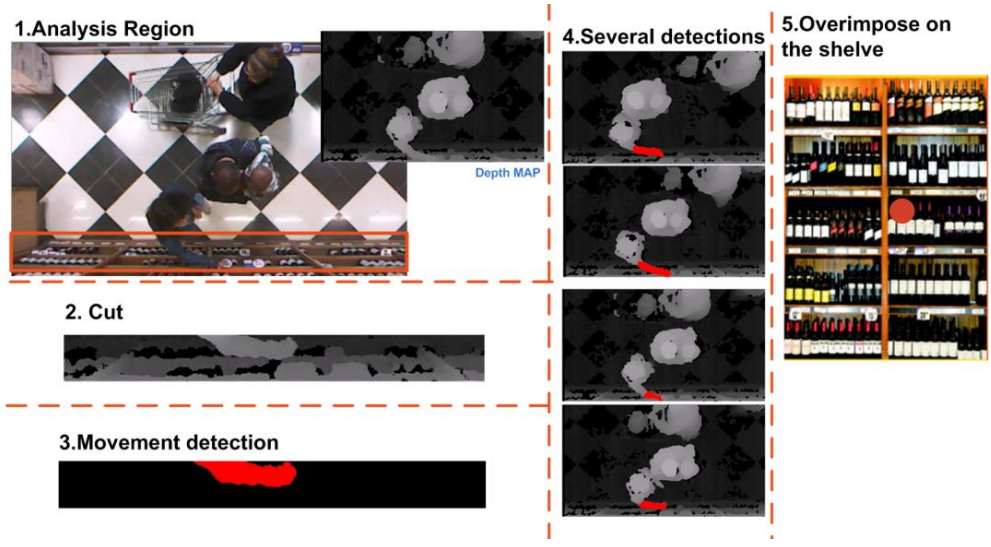


Figure 4 – Step by step pick up detection.

A seudo-code of the tracking and pickups matching algorithm is presented here:

```
Algorithm "Detection, tracking and pickUps"
Input :frame = Actual Video Image frame in T
        activePeople  =List of People detected in previous frame ;
{ Remove far points }
frame = DistanceThreshold(frame, 0,9 meter)
{ detect"people blobs" using the waterFill Algorithm }
List<blob>newB= waterFill(frame)
{ detect interactions only in SHELVES ROI }
List<pickUps>newPU = pickUpsDetection(cut( frame, ROI_SHELVES));
{ Check whether the blob is a new person or an existent one }
For each blob b innewB
    Person p =  activePeople.find(b)
    If p is empty
     Person p =  activePeople.add( new Person(b)){create  new}
    else
     p.path.add(b.position){ updates existent one }

{ check if one pickup corresponds to p }
    For each pu in newPI
      If  connectedPixels(pu, p)
          p.pickUps.add(pu)
end for
{ Evaluate if people is leaving estimating its position innext Frame }
For each people p in activePeople
IfestimatedPosition(p , T + 1) isoutside Frame
activePeople.remove(p)
```

3.3. Classification method

Once an enough number of successive detections are extracted, a behavior can be determined. To perform this analysis, the trajectory of each person is estimated (assuming that the algorithm has properly detected them) and then it is combined with "pick ups" location.

With trajectory length, time in front of camera (active time) and interaction information, people could be classified into 3 categories:

- Walker: person only moving through the place
- Interested: person stopped for certain time
- Purchaser: person who interacts with objects

To distinguish between walker and interested, it is used the active time. As it was empirically estimated, this time should be greater than 5 seconds.

At the same time, analyzing the videos, several different uncommon situations were found. These situations could be detected using particular post-processing techniques and data filters. One of our interests was to find (and discard) people walking too closed that corresponds in general to kids been carried up. This issue could be detected

measuring the distance between two trajectories using DWT [22]. Individuals that are moving close have a minimum distance. For this purpose, trajectories of people that appeared in the same frames were evaluated in pairs and if their distance was below a threshold, one of them was discarded.

Other situation was to detect product suppliers working at peak traffic times (which is not a good commercial practice). These cases were removed from the captured data, filtering information when the time in front shelves and the number of interactions were much higher than the average.

Another situation was congestion cases. Filtering information by the number of people present in the image, it was found that when people stopped to watch a product, it slowed down the normal circulation. This situation resulted from the transit available space and could not be resolved at this time, but served to initiate a study and proposal to improve shelves location In **Fig. 5**, these two cases are shown.

Figure 5–(left) congestion situation (right) product supplier situation

4. SYSTEM ARCHITECTURE

The system is composed of three main modules: the capture, the video processing and the publish one. The capture module connects to devices and generates video streams. The processing module is the most sophisticated one. It takes the image streams and processes them through several layers.

The first layer of this module applies basic image processing techniques to segment objects, detect motion and filter data, as explained in **section 3**. The output of this layer is a set of segmented objects also called "blobs". Combining different features of the blobs, like position, shape or visible time, other tracking and interaction relations could be discovered. With the same data, some false positives could be eliminated. The output of this layer is a list of detected people that have a trajectory and corresponding *pick-ups*. Finally, a characterization layer takes this data, and classifies people and behavior applying clustering techniques. This module is an application that runs on a personal computer.

The layered architecture is shown in Figure 6.

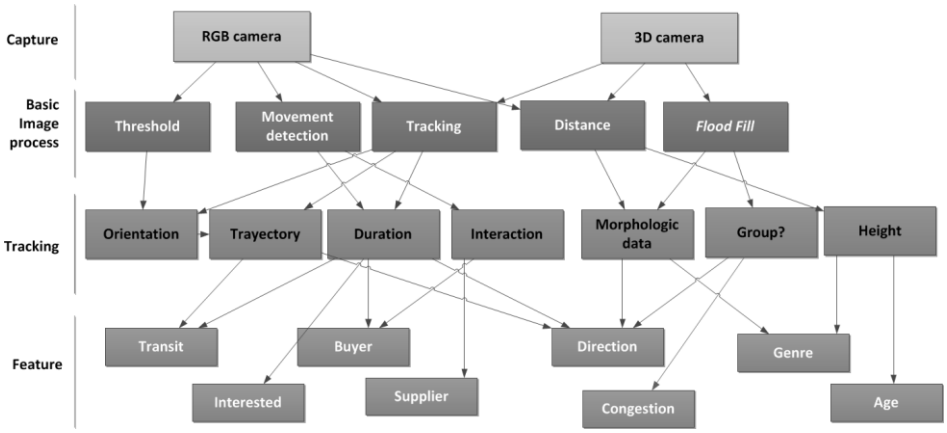


Figure 6 – Processing detection module

Finally, the publish module collects captured data through different days, register the source location (identified by the IP number of the computer), take video samples and store the results in a relational Database Server. This database is accessible from a WEB page that presents information and statistics in a pleasant way. This WEB also offers some facilities to navigate and filter data according to different user permissions and a video player to watch historical records. Captured video are kept in the processing computer (not in the server) and can be watched in real time using a *reverse proxy scheme*. The reverse proxy resolves dynamically where a video is hosted and generates a stream directly between the WEB page and the host. With this scheme, communication bandwidth is greatly optimized, since analyst users are only interested in those parts of the video that have valid detections. The overall architecture is presented in Figure 7

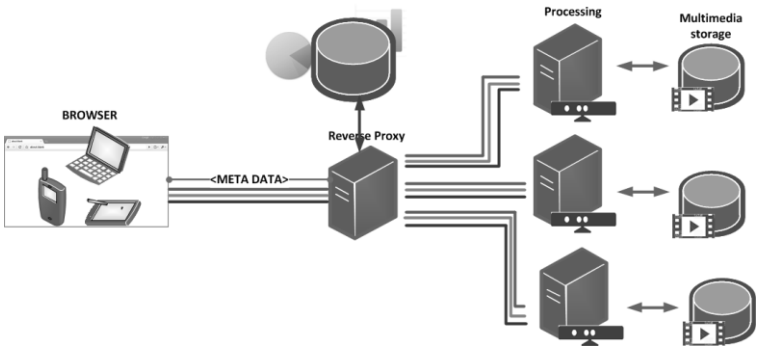


Figure 7 – Overall architecture

5. STUDY CASES

The system was tested mainly on a local supermarket that has a significant transit of more than 1000 people per day. With the marketing managers, it was agreed the

location to be studied, a support structure above the shelves was mounted and the products were manually identified within it. A frontal photo of this location was taken to use as reference. Then, a 100 hours video campaign was made and the results were stored and accessed through an internal network using the proposed architecture. Using the obtained information, we could extract the following indicators:

- Total and daily circulation in the period
- Percentage of movement by hours
- Direction of movement
- Average speed of movement
- Percentage of interested / buyers
- Percentage of bought products classified by brand

Some of these indicators were compared with data procured from the operations responsible.

The first obtained result was the traffic distributed by time bands on weekdays (from Monday to Friday), as shown in Fig. 8. With the system, it was estimated 950 people per day. As was it expected, there is a greater concentration between7p.m. and9p.m..

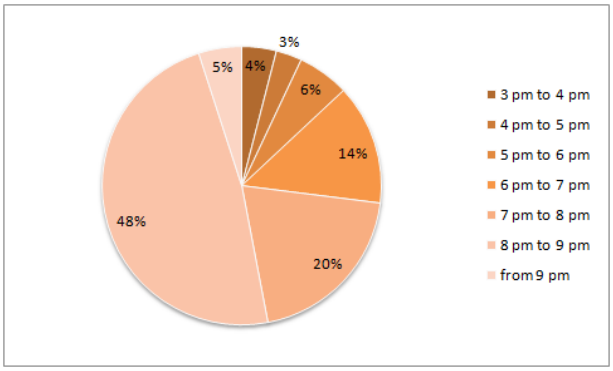


Figure 8 –Amount of people among different hours

The second result was people classification according to the methodology presented. The results were then validated with a specialist who watched and handy accept or reject each of the detection from one day of video (each rejection was a false positives). These results are shown in **Table 2**. Some difference arises, when the same person goes out of the capture region and then immediately returns. The system considers them as two different individuals, when was only one.

Table 2. Classification validation results

	System detection	**False positives**	**Confidence**
Transit	662	21	97%
Interested	227	14	94%
Interaction	61	9	85%

The third indicator extracted was circulation direction. As it was observed, 77% of people come from the left side of the image and 23% from the right. These data

did not have a numerical correspondence (because it was the first time that was measured), but they correspond with the normal market operation.

The final results obtained using this methodology was the spatial distribution of consumer interaction with the product, presented as a *heatmap*. It was noted that for this type of product, when customers interact with a product it was finally acquired. This data was compared with general sales and corresponded in similar proportions. For example, the products (A) and (B) shown in Figure 9 were the most sold products, obtaining a percentage of 38% and 12% in each case.

Figure 9 – Products heatmap

6. CONCLUSIONS

We have presented a customer action classification system that uses computer vision techniques to extract information from video cameras. Data obtained had an interesting impact in understanding how customers move inside a shopping. At the same time it helps to verify if a marketing methodology could be successful. It was possible to implement tracking and people counting algorithms, which could also be corroborated with external systems. Unexpected situations, such as congestion led to location studies that can help to ease the movement of people. Even though, the tracking algorithm fails in some cases, that should be considered.

One of the main advantages of this architecture is flexibility. New capture devices or detection features could be added easily, improving the detection results and removing non valid data. The other one is scalability, as each capture device is managed by a processing unit, the system could be extended adding new units connected through a LAN network. This architecture was used with several HTTP clients but we will continue to test and verify its scalability with more cameras in complex cases. We also are working to integrate information coming from other sources, such as signal data coming from mobile devices; that could be correlated with visual information.

References

[1] Collins, R., Lipton, A. and KanadeT. , "A system for video surveillance and monitoring," In American Nuclear Society 8th Internal Topical Meeting on Robotics and Remote Systems, 1999.

[2] R. Bai..An investigation of novel approaches for optimising retail shelf space allocation. Nottingham, UK : s.n., 2005
[3] Haritaoglu I. and Flickner M. , “Attentive billboards: Towards to video based customer behavior,” In Proc. IEEE Workshop on Applications of Computer Vision, pp. 127-131, Orlando, FL, USA, 2002..
[4] ViCoMo, “Visual Context Modelling,” September 2009, http://www.itea2.org/public/project_leaflets/VICOMO_profile_oct-09.pdf, 2009.
[5] Simpson, L., Taylor,L., O'Rourke, K., Shaw, K. "An Analysis of Consumer Behavior on Black Friday", American International Journal of Contemporary Research Vol. 1 No.1, pp.1-5, 2011.
[6] Senior, A., Brown L., Hampapur A., Shu, C. Zhai Y., Feris, R., Tian, Y. Borger S., and Carlson, C. “Video analytics for retail,” In Proc. IEEE Conference on Advanced Video and Signal-based Surveillance, pp. 423-428, 2007.
[7] Tang Z. and Miao Z. “Fast background subtraction and shadow elimination using improved Gaussian mixture model,” In Proc. IEEE International Workshop on Haptic Audio Visual Environments and their Applications, pp. 38-41, 2007.
[8] Jing, G., Siong, C., and Rajan D. , “Foreground motion detection by difference-based spatial temporal entropy image,” In Proc. IEEE Region 10 Conference (TENCON), vol. A, pp. 379-382, 2004.
[9] Lucas,B. andKanade T. , “An iterative image registration technique with an application to stereo vision, 1991.
[10] Hu W., Tan,T. ,Wang, L. andMaybank S. , “A survey on visual surveillance of object motion and behaviors,” System, Man and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 34(3):334-352, 2004.
[11] Isard, M. and Blake, A. “Condensation - Conditional Density Propagation for Visual Tracking ,” International Journal of Computer Vision, 29(1):5-28, 1998.
[12] Nguyen, N., Venkatesh, S. , West, G. and Bui, H. , “Multiple camera coordination in a surveillance system,” ActaAutomaticaSinica, 2003, 29, (3), pp. 408–421, 2003.
[13] Manavoglu, E. Pavlov, D. and Giles, C. , “Probabilistic User Behavior Models,” Data Mining, ICDM 2003, Third IEEE International Conference on Data Mining, pp. 203-210, 2003.
[14] Mikolajczyk, K., Schmid, C. and Zisserman, A. , “Human detection based on a probabilistic assembly of robust part detectors,” In Proc. European Conference on Computer Vision, volume 3021 of Lecture Notes in Computer Science, pp. 69–81, 2004.
[15] Zhang X., Yan, J. Zhen Lei S., Yi, D. ,Li, S. ,"Water Filling: Unsupervised People Counting via Vertical Kinect Sensor", IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp.215-220, 2012.
[16] Popa, M. Rothlrantz, L., Yang, Z. and Wiggers, P. , "Analysis of Shopping Behavior based on Surveillance System ", IEEE international Conference on Systems Man and Cybernetics (SMC), pp. 2512 – 2519, 2010.
[17] Jagielski, J. “Advanced Reverse Proxy Load Balancing”, Technical report, 2007.
[18] Gervasoni, L, D’Amato, J. Barbuzza, R, Vénere, M. , “Un métodoeficientepara la sustracción de fondo en videos usando GPU”, Revista de MecánicaComputacional(MECOM), vol 34, pp. 4048-4056, 2014.
[19] R. Rajashekar , V. Amarnadh , and M. Bhaskar “Evaluation of stopping criterion in contour tracing algorithms”. Int. J. of Computer Science and Inf. Technologies, 3(3):3888–3894, 2006.
[20] Shopping Behavior Xplained “SBLX” .“Axis cameras watch shopper’s behavior,” http://www.axis.com/files/success_stories/ss_ret_sbxl_36113_en_0907_lo.pdf, 2009.
[21] A. Vacavant , T. Chateau , A. Wilhelm A., and L. Lequièvre A benchmark dataset for outdoor foreground/background extraction. pages 291–300, 2013.
[22] P. Senin: Dynamic time warping algorithm review. Technical Report CSDL-08-04,Department of Information and Computer Sciences, University of Hawaii, Honolulu, Hawaii, 2008.