Multi-Sensor People Counting

Daniel Hernández-Sosa, Modesto Castrillón-Santana, Javier Lorenzo-Navarro *

SIANI
Universidad de Las Palmas de Gran Canaria

Abstract. An accurate estimation of the number of people entering / leaving a controlled area is an interesting capability for automatic surveillance systems. Potential applications where this technology can be applied include those related to security, safety, energy saving or fraud control. In this paper we present a novel configuration of a multi-sensor system combining both visual and range data specially suited for troublesome scenarios such as public transportation. The approach applies probabilistic estimation filters on raw sensor data to create intermediate level hypothesis that are later fused using a certainty-based integration stage. Promising results have been obtained in several tests performed on a realistic test bed scenario under variable lightning conditions.

Keywords: people counting, EKF, MHI, laser sensors

1 Introduction

Automatic surveillance systems are becoming more and more frequent nowadays. People counting constitutes a relevant component of those for many applications. For example, the number of passengers getting in/out of a public transport is necessary for control and management. In pubs and discos the evacuation protocols are designed according to the building capacity and it must not be exceeded. Another example is the presence control for implementing energy saving politics.

Two main technologies have been used to solve the people counting problem: Computer Vision and light beams. On one hand, Computer Vision techniques has been successfully applied to more and more areas in the recent years. This process is favored by the introduction of lower-cost higher-performance hardware and the improvements in the reliability of detection methods. On the other hand, laser sensors have also evolved in the same directions, so smaller and lighter units are available at a reasonable cost vs. precision ratio.

1.1 Computer Vision Methods

In the literature, we can find many examples of Computer Vision based systems with cameras located both in zenithal and non zenithal position. However for some applications where privacy preserving is a crucial matter, the use of vision-based systems with non zenithal cameras is not permitted.

* This work was partially supported by the Spanish MICIIN funds (TIN2008-06068).

Chant et al. [3] proposed a method based on analysing a crowd and making use of mixture of dynamic textures to segment the crowd into different directions; after a perspective correction, some features are computed on each segment and with a Gaussian Process the number of people per segment is obtained. Bozzoli et al. [2] introduced a method for people counting in crowded environments as bus or train gates. The proposal is based on the computation of a running average-like background model applied to edge images in order to avoid the influence of sudden lighting condition changes. Foreground edges are filtered and with the remaining one the optical flow image is computed. Finally each movement vector is assigned to a segment and all the movement vectors assigned to the same segment can be used to estimate the people passing in each direction.

Vision based techniques are well suited for large, wide and open areas, like train platforms or commercial areas besides gates or corridors, provided that lightning conditions are kept under control.

1.2 Range Laser Methods

Katabira et al. [5] proposed a system based on a sensor mounted on the ceiling of a passage. From the range data acquired by the sensor human shapes can be obtained by transforming the data to $X - Z$ plane. The proposed method detects a passing pedestrian when a prominent object is detected.

Mathews and Poigné [8] introduced a system based on a set of passive infrared beacons. The detection of people is done with an Echo State Network which is trained with a set of motion patterns obtained with a simulator.

Light beams based systems have the advantage of privacy preserving, and are best suited for small areas.

1.3 Hybrid Methods

In order to come together the advantages of light beam and vision based systems, some authors have proposed to fusion laser and camera data [9].

Gwang et al. [7] make use of a laser beam as a structured light source. In this way, 3D estimation can be done in an area by means of the integration of consecutive images. When people cross the area, the obtained pattern allows to count the number of people and also the direction of the movement.

Cui et al. [4] describe a method that fuses data from a laser and a visual tracker. The laser module is based on the integration of several laser readings to detect pair of legs and later tracked using a Kalman filter to estimate the position, velocity and acceleration of both feet. A calibrated camera allows to perform visual tracking with color information which feed a mean-shift tracker. Finally, the results of both tracking process are fused with a Bayesian approach.

1.4 The Proposal

In this paper, we propose a fast processing multi-sensor method for counting people getting in/out through a controlled area, using low-cost infrastructure

and preserving privacy. The system is specially well suited for troublesome scenarios with space limitations and changing lightning conditions, such as public transportation applications. Laser and visual based detectors run asynchronously generating hypothesis of crossing people. An upper level combines these hypothesis in order to accept or reject them. The laser process basically extracts relevant peaks from a dynamically generated 3D surface, while the vision process makes use of motion history images to obtain direction and location of people.

The paper is organized as follows: Section 2 gives a brief description of the system. Section 3 presents the results achieved in the experiments. Finally, in the conclusions, some remarks and future work are presented.

2 System Description

The main purpose of our system is to count the number of persons leaving and entering a space. For this work we have considered a specially challenging problem, the monitoring of access to public transportation. In this scenario, an automatic detection system must cope with adverse factors such as severe space limitations and varying lightning conditions. Additionally, the global system cost should be kept reasonably low and subject's privacy should be guaranteed.

The combination of the aforementioned factors has implications on the processing system, as low cost hardware translates into poor data quality and slow acquisition rate. For example, depending on people crossing speed a basic laser sensor can only obtain 3 to 4 scans per individual. Also, height limitation generates important occlusions when a tall person enters/leaves the controlled area, both in a camera or laser data. Besides, due to normally under-illuminated conditions, a standard camera auto-iris is generally wide open, making the depth focus thinner and producing blurring while adjusting to different height people.

The proposed counting people system is composed of a standard webcam and a low cost laser based range sensor. This seems to be an interesting configuration, as lasers provide precise range data but on a small area, while cameras cover a wider area but with a worse signal/noise ratio. Unlike previous works based on fusion of camera and laser readings [4, 9], the range sensor is placed zenithally next to the camera. This configuration is better suited for narrow areas as public transports where the horizontal configuration of the laser is not recommended due to maintenance problems. Additionally, the zenithal location of the camera avoids the privacy matter because faces are not grabbed. The use of low cost sensors allows for a wider economically affordable deployment.

The software architecture is based on a fast pre-attentive visual motion detection and range processing stage, an intermediate data fusion and filtering layer and a final certainty-based decision module. In a previous phase, both laser and camera need to be calibrated, and a region of interest is defined for each sensor. As a result of this calibration process, two coordinate transformation matrices, M_l and M_c, are obtained for laser and camera respectively.

2.1 Visual processing

Two main elements are involved in the visual processing: motion detection and data filtering. The detection uses a motion-energy/motion-history framework to identify module and direction of displacement on images. The data filtering uses an Extended Kalman Filtering (EKF) estimator to integrate motion measures.

Motion detection

Temporal templates have been used as a view-specific representation of movement in the last decade [1]. Image differencing is adequate for some applications, but in our case, as we need to identify if an individual is entering or leaving the room/space, the information contained in a set of consecutive frames is more useful. Motion templates seem to be applicable to this task, and have been used in the past for layering successively the silhouettes obtained after differencing one frame with the previous one.

The motion-energy image (MEI) is a binary cumulative image based on image binary differencing ($D(x, y, t)$) indicating where a motion takes place. Considering a cumulative interval τ, we have the following expression for the MEI image:

$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i).$$

On the other hand, the motion-history image (MHI) indicates also how the motion was produced using each pixel intensity to represent the time elapsed since the motion occurred. In our implementation, we have considered a scalar image, where brighter pixels (max value v_{max}) refer to most recent moving pixels, decaying in intensity (v_{dec} factor) when no difference is present (see Figure 1).

MEI and MHI have been frequently used for gesture recognition [1]. Recently those representations have also been combined with texture descriptors to improve the recognition robustness [6]. However, in our approach, the gestures to recognize are simpler, but different situations can be presented to the sensors due to the various behavioral possibilities that can take place in a door when multiple people are present.

Data filtering

Camera frames are processed to extract motion blobs b_k and filtered out comparing blobs area with a minimum threshold to discard less significant ones ($area(b_k) < area_{min}$). As the camera does not provide depth information, several height possible values (z_1, ..., z_{n_h}) are tested in parallel, using a function that back-projects the blob center coordinates into the corresponding world coordinates via the matrix calibration camera: $b3D_{k,j} = MapXYZ(b_k, M_c, z_j)$ for $j = 1 \dots n_h$.

The set of $b3D_{k,j}$ are used to generate camera-based hypothesis for object trajectories $OTC_{c,j}(t)$, using an EKF framework. The object hypothesis are updated on the basis of the $k{-}th$ detected blob b_k on the current frame, according

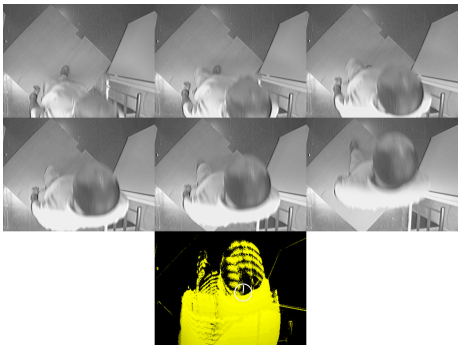

Fig. 1. MHI example.

to the following rule:

$$\begin{cases} \text{if } \min_{i=1\ldots n_c} Dist3D(k,i) < \mu_c & \text{EKF update(all j)}, \\ & OTC_{i,j}(t) \\ \text{otherwise} & n_c = n_c + 1, \\ & \text{EKF init(all j)}, \\ & OTC_{n_c,j}(t) = b3D_{k,j} \end{cases}$$

where $Dist3D(k,i) = \sum_{j=1\ldots n_h} \|b3D_{k,j} - OTC_{i,j}(t-1)\|$, n_c is the number of current active objects and μ_c is a distance threshold.

EKF filters operate on a three state vector representing the objects X and X coordinates and the angle of motion. See image on Figure 2 for an example of blob detection and the corresponding EKF trajectory estimation.

The visual filter keeps integrating data until an object trajectory is detected to intersect the door line, activating then a verification test including the analysis of trajectory length and filter covariance ellipses. On success, the object hypothesis is added to a set H_c of camera-based hypothesis with its spatial and temporal references.

2.2 Laser scan processing

Laser sensors are specially convenient to this problem due to their precision and relative invariance to lightning conditions. In our approach laser readings are integrated over time to generate a kind of 3D topographical surface, $s(x,y,z)$, which exhibits a peak for each person crossing under the sensor ($\nabla(s) = 0$). The $s(x,y,z)$ function is processed to extract relevant peaks which are then fed into a multi-modal tracking filter that keeps laser-based 3D trajectory for hypothetical person objects. These hypothesis, $OTL_i(t)$, are updated on the basis of the $k-th$ detected peak p_k in the current scan according to the following gate-reject/gate-augment rule:

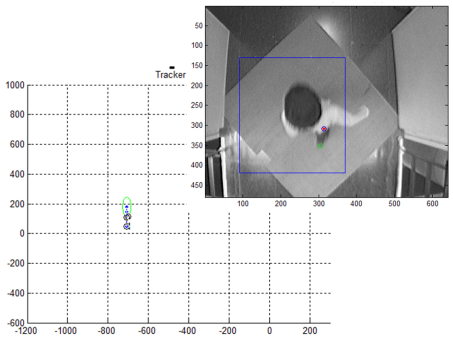

Fig. 2. EKF blobs tracking example.

$$\begin{cases} \text{if } \min_{i=1\ldots n_l} Dist(k,i) < \mu_{rej} & \text{Update } OTL_i(t) \\ \text{if } \min_{i=1\ldots n_l} Dist(k,i) > \mu_{aug} & n_l = n_l + 1, \\ & OTL_{n_l}(t) = p_k \\ \text{otherwise} & discard \end{cases}$$

where $Dist(k,i) = \|ProjXY(p_k, M_l) - ProjXY(OTL_i(t-1), M_l)\|$, and $ProjXY$ is the transformation of the 3D peak coordinates into the XY, n_l is the number of current active object trajectories, μ_{rej} is the gate reject value and μ_{aug} is the gate augment value.

Each time a peak is associated to a trajectory, the area under that peak is computed and integrated to obtain a volume estimation of the object, OV_i, assuming constant velocity.

A trajectory is thus defined in a time interval, starting at its creation $OTL_{i_{ti}}$, and finishing when no peak is associated to the trajectory in the current laser acquisition, $OTL_{i_{tf}}$. Once the trajectory is completed, it is processed to estimate if it could correspond to a crossing person, according to a set of conditions: persistence ($OTL_{i_{tf}} - OTL_{i_{ti}} > t_{min}$), max height ($Max_z(OTL_i) > height_{min}$) and volume ($OV_i > vol_{min}$); where t_{min}, $height_{min}$ and vol_{min} are lower thresholds for trajectory duration, maximum height and volume, respectively. These conditions are defined to try to reduce false positive detections. As a result of this process, a set H_l of laser-based hypothesis about the number and location of people crossing is generated.

2.3 Integration

The high-level heuristic module fuses evidences and sensor data from the detection and filtering layers to generate the final decisions on people crossing events, assigning global certainty values. The two hypothesis sets H_l and H_c are cross

validated to get a more robust estimation of people get in/out counting. Basically, the temporal references (event initial/final time) and spatial references (trajectory coordinates) are used to identify a given object from one set into the other.

The global detection system can label and assign certainty to the crossings depending on they have only range or visual confirmation or both. In general terms, high certainty values are produced when pairs of visual and range data evidences are found to be temporally and spatially coherent. In case of discrepancies, certainty values based on special rules are applied, giving more credibility to the source that has integrated more measures at a reasonable certainty level.

3 Experiments and Results

Several tests have been performed on an experimental setup simulating the conditions of a public transportation conditions, with both visual and range sensors installed at 2.5 meters on a door frame. Data collected include diverse cases of individual and paired in/out door traversing, and simultaneous opposite direction trajectories. Illumination changes have been also introduced artificially switching lights to observe system response. Camera frames were captured from a firewire web-cam in 640x480 format at 30 Hz, while the laser provided 10 Hz scans over a 180 degrees sector with 1 degree angular resolution.

Although sensor configuration was not especially suited due to space and, mainly, height limitations, promising results have been achieved. So, in a sequence of approximately 150 crossing events, around a 90% were correctly detected with the integrated system.

Constant illumination				Illumination changes				Troublesome			
Real		Detected		Real		Detected		Real		Detected	
In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
31	33	29	30	29	28	26	25	12	14	7	6

Table 1. Summary of detection test results.

Table 1 summarizes the results of real vs. detected events obtained in three different scenarios: constant illumination, illumination changes and troublesome. In the first one the illumination was kept constant during the test. In the second one the illumination level was changed artificially several times during the test. The third scenario included a collection of adverse situations such as: crowded group crossing, erratic trajectories, runners, etc.

A small bias can be observed in the table when comparing in and out events. This is due to a vertical sensor misalignment that caused a slightly larger detection area for in events.

Individually considered, laser detection suffers some problems with false positives due to arm movement and poor detection on fast walking people. On the

other hand camera detection showed a lower performance when analyzing tall people crossings and experiences some problems due to auto-iris adjust during illumination changes. The combined detection compensates for these conditions yielding a better global result.

Some situations are clearly not solved by this system configuration, for example children walking close to their parents, people collisions or the use of umbrellas.

4 Conclusions

A low-cost solution to people counting applications in public transportation have been proposed and tested. The combination of range detection and visual detection mechanisms contributes to compensate some specific problems of each method, exhibiting more robust results in getting in/out counting. Regarding more specifically illumination changes, the system is able to discard artificial motion blobs, either on filtering or fusion stages.

Future work includes specific experiments in crowded groups environments and more intensive testing and comparison with range cameras.

References

1. Bobick, A.F., Davis, J.W.: The recognition of human movem. IEEE Transactions on Intelligent Transportation Systems Transactions on Pattern Analysis and Machine Intelligence 23(257 - 267), 3 (March 2001)
2. Bozzoli, M., Cinque, L., Sangineto, E.: A statistical method for people counting in crowded environments. In: 14th International Conference on Image Analysis and Processing (2007)
3. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: Computer Vision and Pattern Recognition. pp. 1 – 7 (2008)
4. Cui, J., Zha, H., Zhao, H., Shibasaki, R.: Multi-modal tracking of people using laser scanners and video camera. Image and Vision Computing 26(2), 240 – 252 (2008)
5. Katabira, K., Nakamura, K., Zhao, H., Shibasaki, R.: A method for counting pedestrians using a laser range scanner. In: 25th Asian Conference on Remote Sensing (ACRS 2004). Thailand (November 22 - 26 2004)
6. Kellokumpu, V., Zhao, G., Pietikinen, M.: Recognition of human actions using texture descriptors. Machine Vision and Applications In press (2010)
7. Lee, G.G., ki Kim, H., Yoon, J.Y., Kim, J.J., Kim, W.Y.: Pedestrian counting using an IR line laser. In: International Conference on Convergence and Hybrid Information Technology 2008 (2008)
8. Mathews, E., Poigné, A.: Evaluation of a "smart" pedestrian counting system based on echo state networks. EURASIP Journal on Embedded Systems 2009, 1–9 (2009)
9. Scheutz, M., McRaven, J., Cserey, G.: Fast, reliable, adaptive, bimodal people tracking for indoor environments. In: Proceedings. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004). vol. 2, pp. 1347– 1352 (2004)