AUTOMATED VISUAL TRAFFIC MONITORING AND SURVEILLANCE THROUGH A NETWORK OF DISTRIBUTED UNITS

A. Koutsia[a], T. Semertzidis[a], K. Dimitropoulos[a], N. Grammalidis[a] and K. Georgouleas[b]

[a] Informatics and Telematics Institute, Centre for Research and Technology Hellas, 1st km Thermi-Panorama road, 57001 Thessaloniki, Greece - (koutsia, theosem, dimitrop, ngramm)@iti.gr
[b] MARAC Electronics, 165 Marias Kiouri & Tripoleos Str, 188 63, Piraeus, Greece - georgouleas@marac.gr

Commission III, WG III/5

KEY WORDS: Computer Vision, Visual Analysis, Fusion, Location Based Services, Calibration, Change Detection, Matching

ABSTRACT:

This work aims to present an intelligent system for tracking moving targets (such as vehicles, persons etc) based on a network of distributed autonomous units that capture and process images from one or more pre-calibrated visual sensors. The proposed system, which has been developed within the framework of TRAVIS (TRAffic VISual monitoring) project, is flexible, scalable and can be applied in a broad field of applications. Two different pilot installations have been installed for initial evaluation and testing, one for traffic control of aircraft parking areas and one for tunnels at highways. Various computer vision techniques which were implemented and tested during the development of the project, are described and analysed. Multiple background extraction and data fusion algorithms are comparatively evaluated.

1. INTRODUCTION

1.1 Relative Work

Traffic control and monitoring using video sensors has drawn increasing attention recently due to the significant advances in the field of computer vision. Many commercial and research systems use video processing, aiming to solve specific problems in road traffic monitoring (Kastrinaki, 2003). An efficient application for monitoring and surveillance from multiple cameras is the Reading People Tracker (Le Bouffant, 2002), which was later used as a base for the development of a system called AVITRACK, which monitors airplane servicing operations (Thirde, 2006). Furthermore, in the FP5 INTERVUSE project, an artificial vision network-based system was developed to monitor the ground traffic at airports (Pavlidou, 2005). The system uses the Autoscope® Solo Wide Area Video Vehicle Detection System which has been successfully deployed worldwide for monitoring and controlling road traffic (Michalopoulos, 1991).

1.2 Motivation and Aims

Robust and accurate detection and tracking of moving objects has always been a complex problem. Especially in the case of outdoor video surveillance systems, the visual tracking problem is particularly challenging due to illumination or background changes, occlusions problems etc. The aim of the TRAVIS project was to determine whether the recent changes in the field of Computer Vision can help overcome these problems and develop a robust traffic surveillance application. The final system is easily adjustable and parameterised, in order to be suitable for diverse applications related to target tracking. Two prototypes have been installed each for a different application:

• Traffic control of aircraft parking areas (APRON). This application focuses more on the graphical display of the ground situation at the APRON. The system calculates the position, velocity and direction of the targets and it classifies them according to their type (car, man, long vehicle etc). Alerts are displayed for dangerous situations, such as speeding. This information can be accessible by the respective employees, even if they are situated in a distant area, with no direct eye-contact to the APRON. A pilot installation of this system took place at "Macedonia" airport of Thessaloniki, Greece.

• Traffic control of tunnels at highways. The focus of this application is on the collection of traffic statistics, such as speed and traffic loads per lane. It can also identify dangerous situations, such as objects falling, animals or traffic jams. These results can be sent to traffic surveillance centres or used to activate road signs/warning lights. This prototype was installed at a highway tunnel at Piraeus Harbour, Athens, Greece.

2. SYSTEM ARCHITECTURE

The proposed system consists of a scalable network of autonomous tracking units (ATUs) that use cameras to capture images, detect moving objects and provide results to a central sensor data fusion server (SDF). The SDF server is responsible for tracking and visualizing moving objects in the scene as well as collecting statistics and providing alerts for dangerous situations. The system provides a choice between two modes, each supporting a different data fusion technique. Grid mode separates the ground plane into cells and fuses neighbouring observations while map fusion mode warps greyscale images of foreground objects in order to fuse them.

The topology of the ATUs network varies in each application depending on the existing infrastructure, geomorphologic facts and bandwidth and cost limitations. The network architecture is based on a wired or wireless TCP/IP connection as illustrated in Figure 1. These topologies can be combined to produce a hybrid network of ATUs. Depending on the available network bandwidth, images captured from specific video sensors may

also be coded and transmitted to the SDF server, to allow inspection by a human observer (e.g. traffic controller).

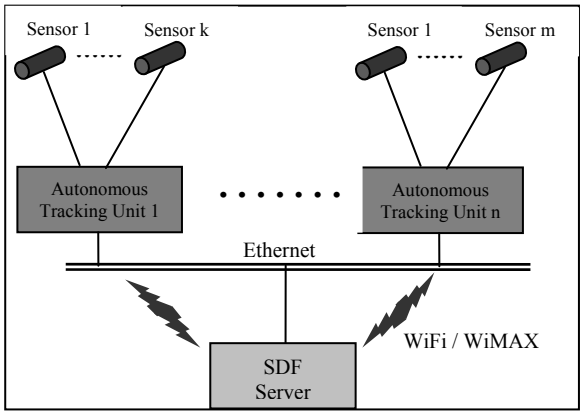

Figure 1: Basic architecture of the proposed system.

3. THE AUTONOMOUS TRACKING UNITS

Each ATU is a powerful processing unit (PC or embedded PC), which periodically obtains frames from one of more video sensors. The video sensors are standard CCTV cameras, not necessarily of high resolution, equipped with a casing appropriate for outdoor use and telephoto lenses for distant observation. They are also static (fixed field of view) and pre-calibrated. Each ATU consists of the following modules:

- Calibration module (off-line unit to calibrate each video sensor). To obtain the exact position of the targets in the real world, the calibration of each camera is required, so that any point can be converted from image coordinates (measured in pixels from the top left corner of the image) to ground coordinates and vice versa. A calibration technique, which is based on a 3x3 homographic transformation and uses both points and lines correspondences, was used (Dimitropoulos, 2005). The observed targets are small with respect to the distance from the video sensors and they are moving on a ground surface, which therefore can be approximated by a plane. For more accurate results, a calibration tool (Figure 2) has been developed. This tool visualises two camera views, one of which is considered the base view according to which the other camera is calibrated. It then allows the user to dynamically choose corresponding points on the two views before it warps them on the ground plane. The user can repeat the procedure until the visual results are considered satisfactory.

- Background extraction and update module. Each ATU of the system can automatically deal with background changes (e.g. grass or trees moving in the wind) or lighting changes (e.g. day, night etc) supporting several robust background extraction algorithms, namely: mixture of Gaussians modelling (KaewTraKulPong, 2001), Bayes algorithm (Liyuan Li, 2003), Lluis-Miralles-Bastidas method (Lluis, 2005) and non-parametric modelling (Elgammal, 2000).

- Foreground segmentation module. Connected component labelling is applied to identify individual foreground objects.

- Blob tracking module (optional). The Multiple Hypothesis Tracker (Cox, 1996) was used, although association and tracking of very fast moving objects could be problematic.

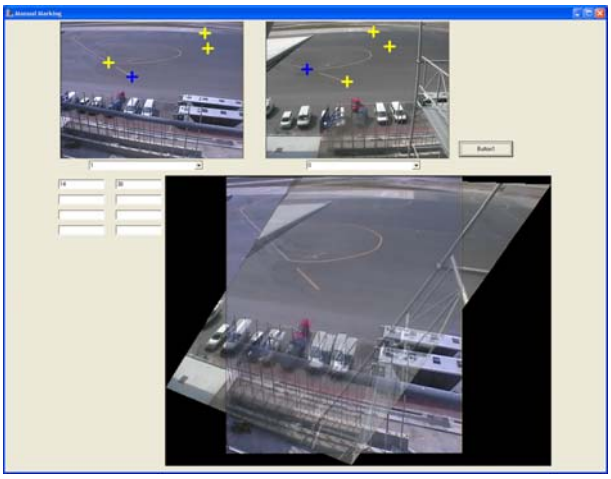

Figure 2: Screenshot from the calibration tool

- Blob classification module. A set of classes of moving objects (e.g. "person", "car" etc) is initially defined for each application. Then, each blob is classified by calculating its membership probability of each class, using a previously trained back-propagation neural network. Specifically, 9 attributes, characteristic of its shape and size, are used as input to a neural network: the two sizes of the major and minor axes of the blob's ellipse and the 7 Hu moments (Hu, 1962) of the blob that are invariant to both rotations and translations. The number of outputs of the neural network equals the predefined number of classes. The class is determined by the maximum output value.

- 3-D observation extraction module. It uses the available camera calibration information to estimate the accurate position of targets in the scene. Since the camera calibration is based on homographies, an estimate for the position (x_w, y_w) of a target in the world coordinates can be directly determined from the centre of each blob. Each observation is also associated with a reliability matrix R, depending on the camera geometry and its position at the camera plane. This matrix is calculated using the calibration information (Borg, 2005):

$$R(x_w, y_w) = J(x_c, y_c)\Lambda J(x_c, y_c)^T \qquad (1)$$

where J = Jacobian matrix of the partial derivatives of the mapping functions between the camera and the world co-ordinate systems,

Λ = measurement covariance at location (x_c, y_c) on the camera plane, which is assumed to be a fixed diagonal matrix.

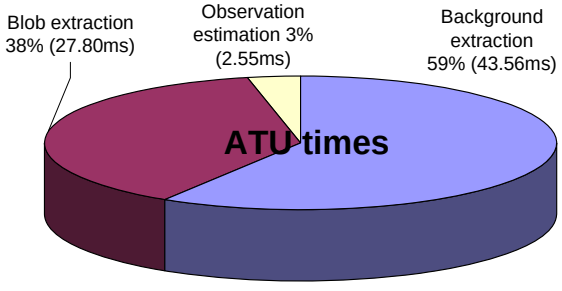

Figure 3: Execution times of ATU modules

Figure 3 shows the absolute and relevant execution times per frame of the basic modules of the ATUs. The particular times have been acquired by applying the non-parametric modelling method, working under grid mode. The background extraction module is the most crucial one as the computational cost of such methods is typically large, causing problems for real-time systems. Therefore, many experiments have been conducted in order to provide both a qualitative and a computational evaluation of these methods.

4. NETWORK COMMUNICATIONS

The final output of each ATU is a small set of parameters (ground coordinates, classification, reliability), which is transmitted to the SDF server through wired or wireless transmission. If the foreground map fusion technique is used, a greyscale image is provided at each polling cycle, indicating the probability for each pixel to belong to the foreground.
All these data are transmitted through wired or wireless IP connection to the server which performs observation fusion and target tracking. TCP protocol is used for transmission of the data from the ATUs to the central server whereas UDP is used for remote controlling of the ATUs. As an indicator, the bandwidth used per ATU when operating under the map fusion mode with a frame rate of 3fps is about 192Kbps (3fps x 8Kbyte/frame).

The system requires frame synchronisation and constant frame rate of all ATUs, which are achieved by using the Network Time Protocol (NTP). The system's clocks synchronise to the central server's clock and a appointment time technique (Litos, 2006) is implemented to ensure that frames from all cameras are captured at the same instant despite network latency.

A secondary system based on media server software streams video on demand to the central server in order to enable human visual monitoring of the scene. As an alternative, compressed motion JPEG images (JPEG 2000) can be used for streaming.

5. SENSOR DATA FUSION SERVER

The SDF Server collects information from all ATUs using a constant polling cycle, produces fused estimates of the position and velocity of each moving target, and tracks these targets using a multi-target tracking algorithm. It also produces a synthetic ground situation display (Figure 4), collects statistical information about the moving targets and provides alerts when specific situations (e.g. accidents) are detected.

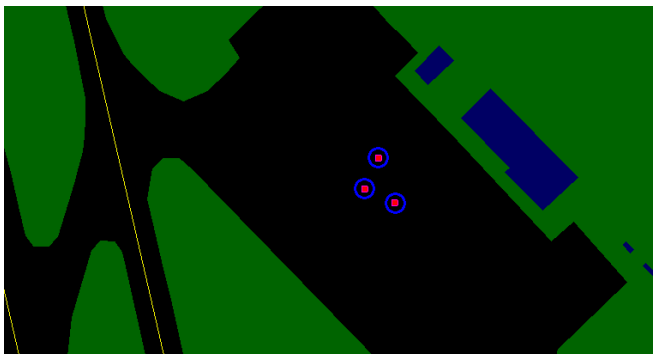

Figure 4: SDF window with 3 targets on the airport APRON

5.1 Data fusion

A target present simultaneously in the field of view of multiple cameras will result in multiple observations due to the fact that the blob centres of the same object in two different cameras correspond to close but different 3-D points. Two techniques are proposed for grouping together all the observations that correspond to the same target:

5.1.1. Grid-based fusion

A grid that separates the overlap area (in world coordinates) in cells is defined. Optimal values for the cell size are determined considering the application requirements (e.g. maximum distance between vehicles). Each observation is assigned two index values (i_x, i_y) that indicate its position on the grid:

$$(i_x, i_y) = ([x_w - x_s] \bmod c, [y_w - y_s] \bmod c) \quad (2)$$

where x_s, y_s = world coordinates of the top left corner of the overlap area
x_w, y_w = world coordinates of the camera level observation
c = cell size

Observations belonging to the same cell or to neighbouring cells are grouped together to a single fused observation.

To implement this technique the grid is expressed as a binary image: cells that have at least one assigned observation are represented by a white pixel, while those with no observations are represented by a black pixel. A connected component labelling algorithm is then used to identify blobs in this image, each corresponding to a single moving target.

Fused observations are produced by averaging the parameters of the observations that belong to each group. More specifically, each fused observation consists of an estimated position of the world coordinates, an uncertainty matrix as well as a classification probability matrix.

The position and uncertainty matrices ($\mathbf{Z}$, $\mathbf{R}$) of the fused observation are given by the following equations:

$$\mathbf{R} = \left(\sum_{n=1}^{N} \mathbf{R}_n^{-1}\right)^{-1} \quad (3)$$

$$\mathbf{Z} = \mathbf{R} \sum_{n=1}^{N} \mathbf{R}_n^{-1} \mathbf{Z}_n$$

where $\mathbf{Z}_n$ = the position (in world coordinates) of the n-th observation in a group of N
$\mathbf{R}_n$ = uncertainty matrix of the n-th observation in a group of N

To calculate the average classification vector, the uncertainty of each observation is taken into account. In this case the larger of

the two axes of the observation uncertainty ellipse is used to specify a weight for the observation in the classification averaging. Depending on the magnitude of this metric, a corresponding weight is assigned to each observation. The average classification vector is then calculated by:

$$\mathbf{c} = \frac{\sum_{n=1}^{N} w_n \mathbf{c}_n}{\sum_{n=1}^{N} w_n} \quad (4)$$

These parameters ($\mathbf{Z}$, $\mathbf{R}$, $\mathbf{c}$) of each fused observation comprise the input for the tracking unit.

5.1.2. Foreground map fusion

In this technique, each ATU provides the SDF server with one greyscale image per polling cycle, indicating the probability for each pixel to belong to the foreground. The SDF server fuses these maps together by warping them to the ground plane and multiplying them (Khan, 2006). The fused observations are then generated from these fused maps using connected component analysis and classification information is computed as in the ATU's blob classification module. Although this technique has increased computational and network bandwidth requirements, when compared to grid-based fusion, it can very robustly resolve occlusions between multiple views.

5.2 Multiple Target Tracking

The tracking unit is based on the Multiple Hypothesis Tracking (MHT) algorithm, which starts tentative tracks on all observations and uses subsequent data to determine which of these newly initiated tracks are valid. Specifically, MHT (Blackman, 1999) is a deferred decision logic algorithm in which alternative data association hypotheses are formed whenever there are observation-to-track conflict situations. Then, rather than combining these hypotheses, the hypotheses are propagated in anticipation that subsequent data will resolve the uncertainty. Generally, hypotheses are collections of compatible tracks. Tracks are defined to be incompatible if they share one or more common observation. MHT is a statistical data association algorithm that integrates the capabilities of:

- Track Initiation: Automatic creation of new tracks as new targets are detected.
- Track Termination: Automatic termination of a track when the target is no longer visible for an extended period of time.
- Track Continuation: Continuation of a track over several frames in the absence of observations.
- Explicit Modelling of Spurious Observations
- Explicit Modelling of Uniqueness Constraints: An observation may only be assigned to a single track at each polling cycle and vice-versa.

Specifically, the tracking unit was based on a fast implementation of the MHT algorithm (Cox, 1996). A 2-D Kalman filter was used to track each target and additional gating computations are performed to discard observation – track pairs. More specifically, a "gate" region is defined around each target at each frame and only observations falling within this region are possible candidates to update the specific track. The accurate modelling of the target motion is very difficult, since a target may stop, move, accelerate, etc. Since only position measurements are available, a simple four-state (position and velocity along each axes) CV (constant velocity) target motion model in which the target acceleration is modelled as white noise provides satisfactory results.

Figure 5 shows the absolute and relevant values of the execution times per frame of the SDF server modules. The particular times have been acquired by working under the map fusion mode. The data fusion module appears to be the most time consuming one.

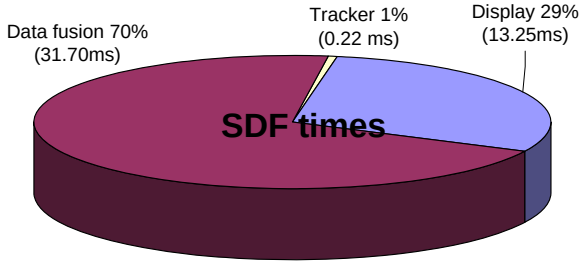

Figure 5: Execution times of SDF modules

6. EXPERIMENTAL RESULTS

In this section, experimental results that concern the most crucial and computationally expensive modules of the ATU and SDF software are presented and discussed. These modules have been identified as the background extraction module for the ATUs and the data fusion module for the SDF server.

For the purposes of deciding on the most appropriate background extraction technique for the specific applications, tests have been run on various sequences. The masks shown on Figure 6 are obtained from the prototype system installation at "Macedonia" airport in Thessaloniki. Figure 6 (a) shows the original image, while in Figure 6 (b) the three moving objects that need to be detected by the background extraction methods are marked with red circles. As seen in Figure 6 (c) the objects are detected with the mixture of Gaussians method, although the shape of the masks is distorted due to shadows. The results of the Bayes algorithm are shown in Figure 6 (d). This method fails to detect slowly moving objects like the one on the left of the image. The Lluis et al method shown in Figure 6 (e) produces masks with low level of connectivity, which are not suitable for the following image processing steps. Finally the non-parametric modelling method (in Figure 6 (f)) yields very accurate results, while coping well with shadows, as it incorporates an additional post processing step of shadow removal.

Another crucial issue when deciding on the most appropriate background extraction algorithm is its execution time. To evaluate the computation complexity, all four methods were applied on three sequences of different resolutions (320x740px, 640x480px, 768x576px). The execution times per frame for each of the four methods and three sequences are presented on Figure 7. An Intel Pentium 4 3.2GHz with 1GB of RAM running on Windows XP Pro was used and all algorithms were implemented in C++ using the open source library OpenCV. Taking into consideration both the qualitative results and the computational complexity of background extraction methods, the non-parametric modelling emerges as the one having the best trade-off between results quality and execution times.

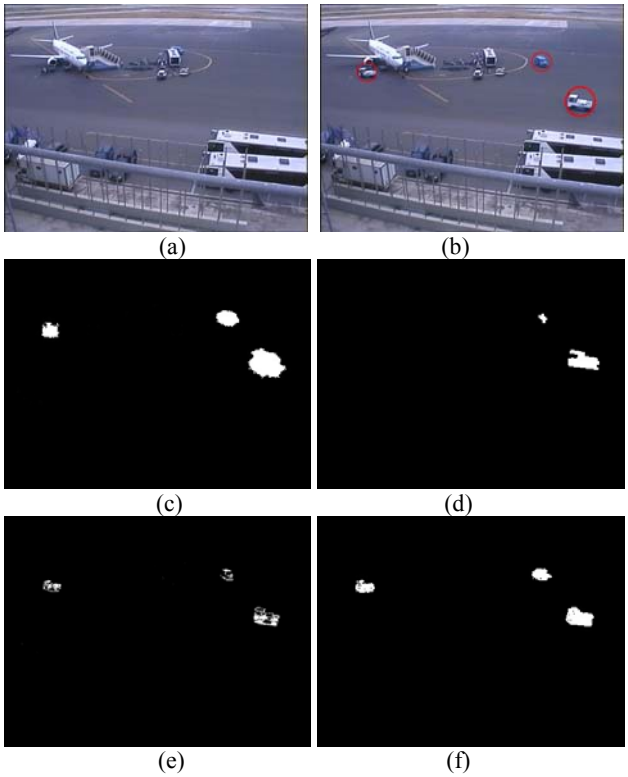

Figure 6: a) Original image b) Moving objects c) Mixture of Gaussians mask d) Bayes algorithm mask e) Lluis-Miralles-Bastidas mask f) Non-parametric model mask.

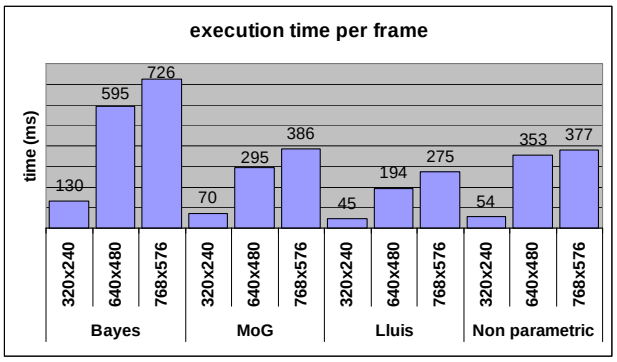

Figure 7: Background extraction methods execution times

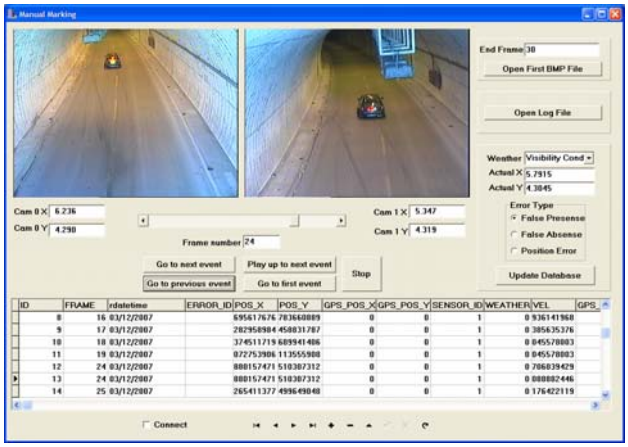

Figure 8: Screenshot from the Manual Marking tool

In order to analyse the results acquired by the system, ground truth data had to be made available. For the airport prototype, tests have been conducted using cars equipped with GPS devices. For the tunnel prototype on the other hand, the ground truth data was collected by viewing image sequences through a specially developed manual marking tool (Figure 8), which allows the user to specify the correct position of the moving targets. The ground truth data has been correlated with the system results and then inserted into a database, along with other information such as weather conditions. A test analysis tool has also been implemented in order to display various statistics that can be acquired by querying the database.

This test analysis tool was used in order to provide qualitative comparison of the two operation modes of the system, the grid and map fusion modes. A very crucial statistic that is suitable for this kind of evaluation is the absence error. Its significance lays on the fact that a high value of this statistic means that the system is prone to be lead to wrong conclusions with severe consequences, such as failing to identify an emergency situation. As it can be seen on Figure 9 the absence error appears more rarely than the other two types of errors (presence and position) for both modes. Especially in the map fusion mode, this statistic is even lower, achieving half the value of the one acquired from the grid mode.

Although the map fusion method appears to provide more accurate results, there are other areas where the grid mode shows better performance, such as the utilisation of less bandwidth. The volume of the data transmitted by each ATU at every frame circle for grid fusion as measured for the prototype applications was 1Kbyte/frame while for the map fusion mode it was measured at 8Kbyte/frame.

Another characteristic of the two methods that is worth mentioning is the execution times they achieve, both for the ATUs and the SDF. As shown in Figure 10, the grid mode is more intensive for the ATUs while the map fusion mode evokes bigger execution times at the SDF server. In other words, the choice of operation mode is a complex decision that should be based on several factors, such as the network capabilities, the available computation power of the SDF server unit and the total number of ATUs used on the particular configuration of the system.

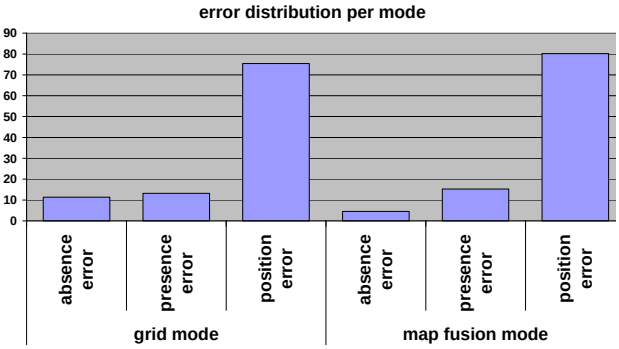

Figure 9: Percentage of error distribution per mode

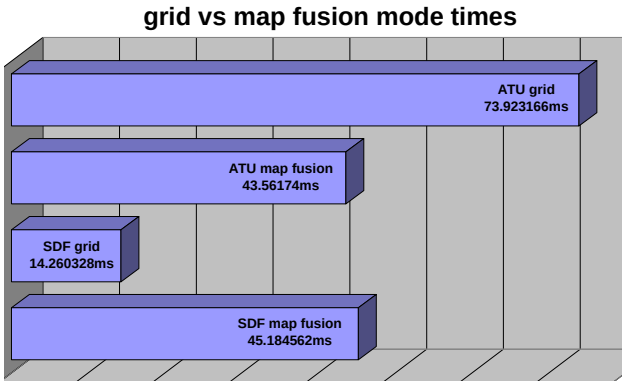

Figure 10: ATU and SDF execution times for both modes

7. CONCLUSIONS

This paper presented an automated solution for visual traffic monitoring based on a network of distributed tracking units. The system can be easily adjusted and parameterised in order to be used in several traffic monitoring applications, as it was built based on results acquired from two diverse pilot installations. The first prototype, installed at an airport APRON, was using an outdoor scene with large field of view while the second prototype, installed in a highway tunnel, was using an indoors scene with smaller distances and more occlusions. The results presented were focused on the two most important modules of the system, the background extraction method and the data fusion technique. After both qualitative and quantitative evaluation of multiple alternatives, the non-parametric modelling method was chosen as the best solution for the system, regarding the background extraction module. On the other hand, both the data fusion techniques tested showed satisfying behaviour under different situations and the final choice between the two should depend on the specific application demands and infrastructure. An interesting future extension is to take advantage of the low bandwidth output of the SDF server in order to create a 3D synthetic representation of the scene under surveillance, which could be rendered at remote 3D displays.

REFERENCES

Blackman, S., Popoli, R., 1999. *Design and analysis of modern tracking systems*. Artech House, Boston, USA.

Borg, M., Thirde, D., Ferryman, J., Fusier, F., Valentin, V., Brémond, F., Thonnat, M., Aguilera, J., Kampel, M., 2005. Visual Surveillance for Aircraft Activity Monitoring. *VS-PETS 2005*, Beijing, China.

Cox, J., Hingorani, S.L., 1996. An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and its Evaluation for the Purpose of Visual Tracking. *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 18, pp. 138-150.

Dimitropoulos, K., Grammalidis, N., Simitopoulos, D., Pavlidou, N., Strintzis, M., 2005. Aircraft Detection and Tracking using Intelligent Cameras, *IEEE International Conference on Image Processing*, Genova, Italy, pp. 594-597.

Elgammal, A., Harwood, D., Davis, L., 2000. Non-parametric Model for Background Subtraction.Computer Vision. *ECCV 2000*.

Hu, M-K., 1962. Visual pattern recognition by moment invariants. *IRE Trans. on Information Theory*, IT-8:pp, 179-187.

KaewTraKulPong, P., Bowden, R., 2001. An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection. In *2nd European Workshop on Advanced Video-based Surveillance Systems*, Kingston, UK.

Kastrinaki, V., Zervakis, M., Kalaitzakis, K., 2003. A survey of video processing techniques for traffic applications. *Image Vision Computing*, Volume: 21,Issue: 4, pp. 359 – 381.

Khan, S., Shah, M., 2006. A multiview approach to tracking people in crowded scenes using a planar homography constraint. *9th European Conference on Computer Vision*, Graz, Austria.

Le Bouffant, T., Siebel, N. T., Cook, S., Maybank, S., 2002. Reading People Tracker Reference Manual (Version 1.12), Technical Report No. RUCS/2002/TR/11/001/A, Department of Computer Science, University of Reading.

Litos, G., Zabulis, X., Triantafyllidis, G.A.,2006. Synchronous Image Acquisition based on Network Synchronization, *IEEE Workshop on Three-Dimensional Cinematography*.

Liyuan Li, Weimin Huang, Irene Y.H. Gu, Qi Tian, 2003. Foreground Object Detection from Videos Containing Complex Background. In *International Multimedia Conference*.
Lluis, J., Miralles, X., Bastidas, O., 2005. Reliable Real-Time Foreground Detection for Video Surveillance Application. In *VSSN'05*.

Michalopoulos, G., 1991. Vehicle Detection Video Through Image Processing: The Autoscope System. *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1.

Pavlidou, N., Grammalidis, N., Dimitropoulos, K., Simitopoulos, D., Gilbert, A., Piazza, E., Herrlich, C., Heidger, R., Strintzis, M., 2005. Using Intelligent Digital Cameras to Monitor Aerodrome Surface Traffic. *IEEE Intelligent Systems.* Vol. 20, No. 3, pp.76-81.

Thirde, D., Borg, M., Ferryman, J., Fusier, F., Valentin, V., Bremond, F., Thonnat, M., 2006. A Real-Time Scene Understanding System for Airport Apron Monitoring. In *Proceedings of the Fourth IEEE international Conference on Computer Vision Systems*.

ACKNOWLEDGEMETS

This work was supported by the General Secretariat of Research and Technology Hellas under the InfoSoc "TRAVIS: Traffic VISual monitoring" project and the EC under the FP6 IST Network of Excellence: "3DTV-Integrated Three-Dimensional Television - Capture, Transmission, and Display" (contract FP6-511568).