# FLEXIBLE TEST-BED FOR UNUSUAL BEHAVIOR DETECTION

István Petrás[1]
Csaba Beleznai[2]
Yiğithan Dedeoğlu[3]
Montse Pardàs[5]

Levente Kovács[1]
Zoltán Szlávik[1]
László Havasi[1]
Tamás Szirányi[1]

B. Uğur Töreyin[4]
Uğur Güdükbay[3]
A. Enis Çetin[4]
Cristian Canton-Ferrer[5]

## ABSTRACT

Visual surveillance and activity analysis is an active research field of computer vision. As a result, there are several different algorithms produced for this purpose. To obtain more robust systems it is desirable to integrate the different algorithms. To help achieve this goal, we propose a flexible, distributed software collaboration framework and present a prototype system for automatic event analysis.

## Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: - Scene Analysis, *Motion, Tracking*, I.5.5 [**Pattern Recognition**] - Implementation, *Special architectures*

## General Terms

Algorithms, Theory.

## Keywords

visual surveillance, anomaly detection, distributed system, event detection,

## 1. INTRODUCTION

Visual surveillance and activity analysis has attained great interest in the field of computer vision research [1][2]. Several algorithm libraries are available on-line (open-source or proprietary), however their integration into a complex system is hindered by the inhomogeneity of the implementation language, format, processing speed, etc. The aim of this work is to produce a flexible system for activity analysis. We provide a transparent and distributed architecture for easy integration of third party modules into a common framework to facilitate easier research collaboration and evaluation. The setup is hierarchical thus helping the scalability of the whole framework. The actual implementation integrates diverse algorithms forming a test-bed for unusual activity detection. Various complex surveillance related algorithms, such as human and body action, tracking and motion activity algorithms are integrated into one system. In the case of detecting unusual motion occurrences, we refer to the term *unusual* in statistical sense.

## 2. SYSTEM ARCHITECTURE

The architecture according to the current trend and software tools is as flexible as possible. The modules can be distributed over the network (either LAN or WAN); they are organized into a hierarchical structure. The structure can be separated into four main entities: a) the clients, b) the server (optionally including the web server) c) the controller and d) the communication interface embedded into the user module (see Figure 1). Each component operates autonomously communicating through RPC requests over TCP/IP.

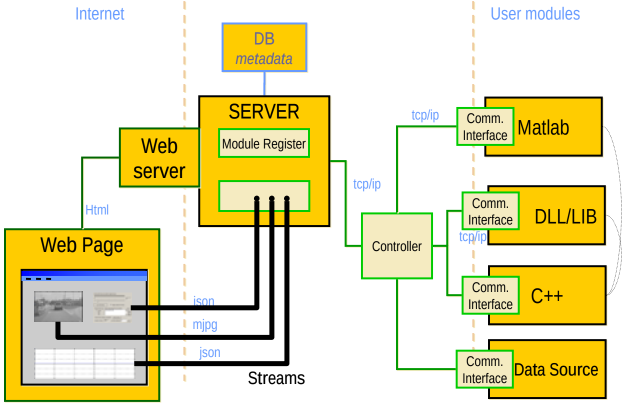


**Figure 1: System architecture**

a) The web-based control interface of the server (Figure 2) gives the user transparent access to the control of the modules and can display their current status and output. It is written in Java and XHTML/ActionScript.

b) The server is the central element of the system; it sits on the top of the hierarchy. It delegates the tasks to the modules and coordinates the execution of the different modules: starts and stops modules and sets their parameters. The server can have more than one controller connection.

c) The controller is a mid-level entity. Its role is to serve as a gateway between the modules on the local network. One controller can have several modules connected to it.

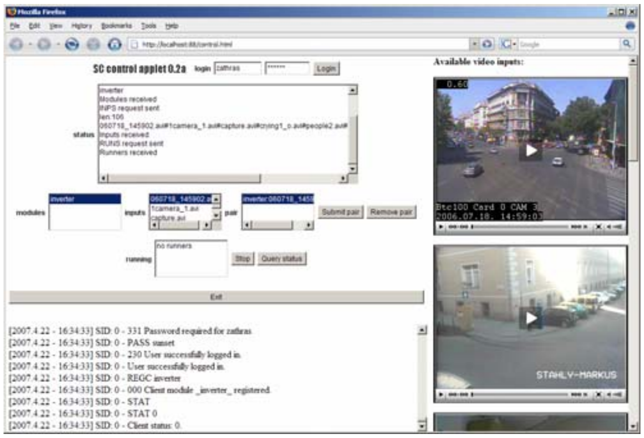
d) The communication interface seamlessly integrates into the user module. It is implemented as a C++ class. It requires as little changes to the user modules as possible and it is designed to be easily integrated into third parties' existing applications. Through the interface the modules can transparently exchange images frames and other data.

One advantage of the architecture is that there is no need to

*CIVR'07*, July 9-11, 2007, Amsterdam, The Netherlands.

[1] MTA-SZTAKI, H-1111 Budapest Kend 13-17. +36 1 2797300, petras@sztaki.hu
[2] Advanced Computer Vision GmbH - ACV, Vienna, Austria
[3] Bilkent University, Department of Computer Engineering
[4] Department of Electrical and Electronics Engineering, 06800, Bilkent, Ankara, Turkey
[5] Department of Signal Theory and Communications, Universitat Politecnica de Catalunya C/Jordi Girona 1-3, D-5, 08034 Barcelona, Spain

publicly provide any proprietary source code, yet it is still possible to integrate heterogeneous modules through TCP/IP. Third parties' IP rights can be easily protected as they can run their own modules on their own servers; they only need to incorporate the communication interface classes we provide.



login: zathras

```
inverter
Modules received
INPS request sent
len:106
060718_145902.avi#1camera_1.avi#capture.avi#crying1_o.avi#people2.avi#
inputs received
RUNS request sent
Runners received
```

modules: inverter

inputs: 060718_145902.a… / 1camera_1.avi / capture.avi

pair: inverter:060718_145…

running: no runners

```
[2007.4.22 - 16:34:33] SID: 0 - 331 Password required for zathras
[2007.4.22 - 16:34:33] SID: 0 - PASS sunset
[2007.4.22 - 16:34:33] SID: 0 - 230 User successfully logged in.
[2007.4.22 - 16:34:33] SID: 0 - User successfully logged in.
[2007.4.22 - 16:34:33] SID: 0 - REGC inverter
[2007.4.22 - 16:34:33] SID: 0 - 000 Client module _inverter_ registered.
[2007.4.22 - 16:34:33] SID: 0 - STAT
[2007.4.22 - 16:34:33] SID: 0 - STAT 0
[2007.4.22 - 16:34:33] SID: 0 - Client status: 0.
```

**Figure 2: Simple web interface**

# 3. ACTUAL MODULES

In the following we describe the actual modules in the system and their background.

## 3.1 Body actions module

### Human model and motion based unusual event detection

In order to achieve a simple motion representation, [6] introduced the concept of Motion History Image (MHI) and Motion Energy Image (MEI). This representation has been recently used for monocular gait recognition tasks [7] and activity modeling [9]. We have extended this formulation to represent view-independent 3D motion [5]. A simple ellipsoid body model was fit to the incoming 3D data to capture in which body part the gesture occurs thus increasing the recognition ratio of the overall system and generating a more informative classification output. In this module, we use the same approach for monocular sequences. The system is based on three sub-modules:

**2D body model.** The inputs to this module are the foreground regions previously extracted by the background/foreground separation module. In order to extract a set of features describing the body of a person that performs an action, a geometrical configuration of human body must be considered. Since the aim of our research is to increase robustness of gesture classification by embedding human body configuration information in our data analysis loop while keeping real-time performance, a simple elliptic model of human body has been adopted. For each connected component classified as foreground we fit an ellipse using statistic moment analysis, obtaining the spatial mean and covariance matrix of the pixels belonging to the connected component. The obtained parameters, together with the horizontal and vertical projection histograms of the silhouettes of the blobs extracted are first used to identify if the blob belongs to a single person or not (that is, it corresponds to a group of people or to any other object such as a car or truck). Only if the blob is classified as a person it will be further processed.

**Motion modeling.** The binary Motion Energy Image (MEI) Eτ (x; t) is defined as:

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

where $\Omega_D$(x; t) is the binary data set indicating regions of motion. This measure captures the locations where there is motion in the last τ frames. Motion detection captured in $\Omega_D$ (x; t) can be coarsely estimated by a simple forward differentiation among pixel frames, still leading to satisfactory results while preserving a reduced computational complexity. It should be noted that τ is a crucial parameter in defining the temporal extent of the actions.

To represent the temporal evolution of the motion, we define the Motion History Image (MHI) where each of the pixel intensities is a function of the temporal history of the motion at that location. Formally,

$$H_\tau(x,t)\begin{cases}\tau & if \quad \Omega_D \quad (x,t)=1\\ \max[0,H_\tau(x,t-1)-1] & otherwise\end{cases}$$

This particular choice of temporal projection operator has the advantage that computation is recursive thus being a good representation for a real-time system. An example of MHI is shown in Figure 3.

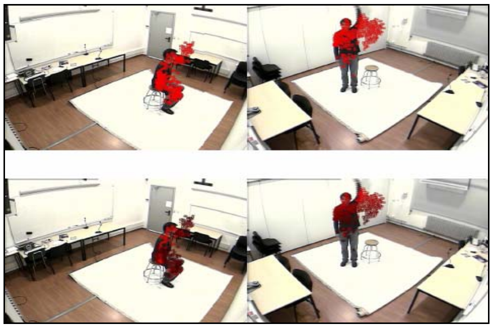


**Figure 3: Examples of motion descriptors: First row MEI and second row MHI**

**Unusual event detection.** Data produced by the body and motion analysis modules is processed in order to extract a vector of features for classification. Informative features derived from the analyzed data (MHI and MEI) are required. Statistical moments invariant to scaling, translation, rotation and affine mappings have been used [8]. For each data set (MHI and MEI), 5 invariant moment-based features are computed. Information from the elliptic body model can be used to generate additional features. We use two features describing the relative amount of motion pixels located in the upper and lower body part, defined through the elliptic model. Thus, we construct a 12-dimensional feature vector. For each scenario, this feature vector is trained for the usual events (people walking and people standing for instance) using a mixture of Gaussians probability model. The detection of unusual events is based on a classification of each feature vector as belonging to this model or not.

## 3.2 Pedestrian and tracking module

**Non-parametric clustering for object detection.** Fast mean shift-based clustering in 2D digital images is introduced using integral images. The fast clustering step is used to delineate objects directly in a difference image obtained by a standard adaptive background subtraction technique in an automated visual surveillance system [10].

The clustering step requires a single parameter in form of an object (kernel) size model *H(x)* and does not rely on any sensitive parameters such as thresholding.

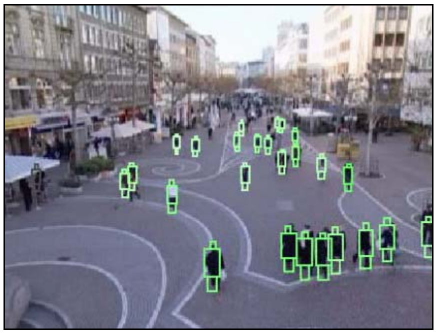


**Figure 4: Human detection results**

**Occlusion handling for interacting targets.** A novel occlusion handling scheme is implemented, which significantly improves the tracking performance even in the presence of a large overlap between objects [11]. The optimal spatial arrangement, i.e. *configuration* of occluding targets is determined by searching for the maximum joint-likelihood estimate in the space of possible object configurations. The configurations are efficiently evaluated using integral image-based computations. The search employs a sampling scheme relying on the mean shift procedure and on priors with respect to the number and size of involved targets (Figure 4).

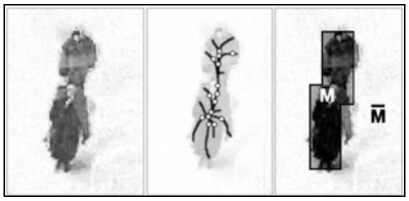


**Figure 5: Computation of the maximum likelihood object configuration using fast integration of the difference image (difference image is shown inverted)**

**Kernel-based tracking using motion features for multiple targets** (Figure 6)**.** A kernel-based fast tracking algorithm [12][13] is applied to the track density maxima in a difference image. The principal advantages of this tracking strategy are: (1) the data association problem is solved implicitly, since the mode seeking procedure is guided to the nearby mode along the steepest density gradient (Figure 5); (2) it represents a simple and computationally efficient technique, because only a few fast mean shift iterations are sufficient to re-detect the object. The mode seeking process is complemented with a linear motion model.

**Video data set and evaluation framework.** We proposed a motion detection and tracking evaluation framework using semi and full-synthetic video sequences under controlled variation of selected parameters [14]. Objective measures assessing the motion detection and tracking quality were proposed and various algorithms have been compared [15].

An annotation framework has been developed within the MUSCLE network of excellence and two datasets with annotations have been made publicly available.

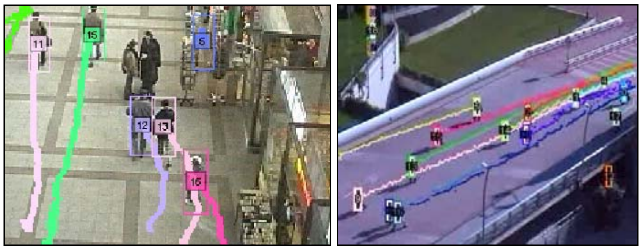


**Figure 6: Fast kernel-based tracking of multiple targets using motion information.**

## 3.3 Multimodal human actions module

Recently, intelligent video analysis systems capable of detecting humans, cars etc were developed. Such systems mostly use HMMs or SVMs to reach decisions. They detect important events but they also produce false alarms. It is possible to take advantage of other low cost sensors including audio to reduce the number of false alarms [1]. Most video recording systems have the capability of recording audio as well. Analysis of audio for intelligent information extraction is a relatively new area. Automatic detection of broken glass sounds, car crash sounds, screams, increasing sound level at the background are indicators of important events. By combining the information coming from the audio channel with the information from the video channels, reliable surveillance systems can be built.

In this module (Figure 7), a surveillance subsystem that utilizes both video and audio to detect fight among people at unattended places were developed. First, moving objects in video are segmented from the scene background by using an adaptive background subtraction algorithm and then segmented objects are classified into groups like human and human group using a silhouette based classification method [4]. By analyzing the motion of the human groups and at the same time detecting screams or increasing sound in audio a decision is given to detect fight.

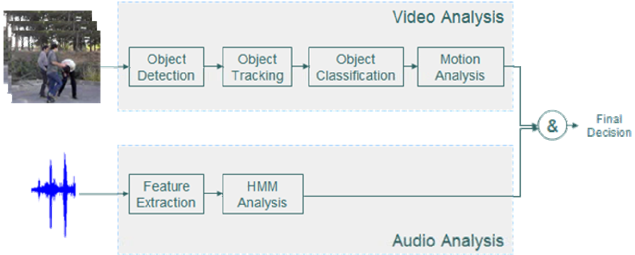


**Figure 7: Block diagram of the body actions module**

## 3.4 Unusual motion pattern module

Intelligent visual surveillance is an increasingly important part of computer vision research. One of the most important goals of visual surveillance systems is to analyze the activity of the observed objects in order to detect anomalies, predict future behaviors, or predict potential unusual events before they occur. There have been a lot of approaches to model the activity of dynamic scenes. Analysis of motion patterns is an effective approach for learning the observed activity. For the most of the time, objects in the scene do not move randomly. They usually follow well-defined motion patterns. Knowledge of usual motion patterns can be used to detect anomalous motion patterns of objects. Current systems mainly base their analysis of motion patterns on a predefined classification of tracked data [16] or of optical flow patterns [17].

The module consists of three processing sub-modules. First, the input image sequence is filtered to remove noise and enhance significant edges. Then the EMD (elementary motion detection [18][19]) module estimates the local motion vectors in every pixel which are fed into the GMM (Gaussian mixture model) sub-module. The GMM module learns the usual motion patterns and evaluates whether the input pixel values are from the learned sets or not. Finally, the Decision module evaluates the output of the GMM sub-module and produces an alarm signal if an unusual motion pattern was detected (Figure 8).

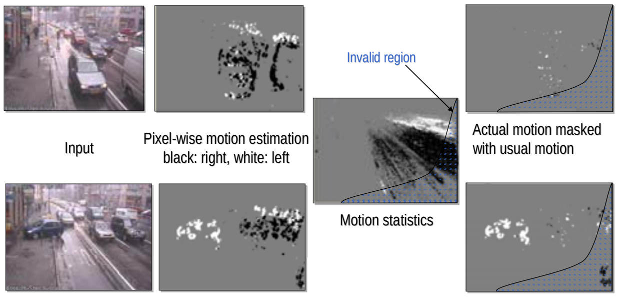


**Figure 8: Motion pattern based anomaly detection.**

### 3.5 Fusion and evaluation module

The purpose of this module is to fuse the output of the modules and trigger an alarm. This alarm signal is passed to the server and then displayed in the web-based client.

## 4. ACKNOWLEDGMENTS

We kindly acknowledge the financial support of the EU - MUSCLE Network of Excellence. Part of this work has been carried out within the K plus Competence Center Advanced Computer Vision. This work was partially funded from the K plus Program.

## 5. REFERENCES

[1] W. Hu et al., "A Survey on Visual Surveillance of Object Motion and Behaviors," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 3, 2004, pp. 334–352.

[2] Hilary Buxton: Learning and understanding dynamic scene activity: a review. Image Vision Comput. Vol. 21, no. 1: pp: 125-136, 2003

[3] B. Uğur Töreyin, Yiğithan Dedeoğlu, A. Enis Çetin, HMM Based Falling Person Detection Using Both Audio and Video, IEEE Int. Workshop on Human-Computer Interaction, Beijing, China, (in conjunction with ICCV 2005), *Lecture Notes in Computer Science*, vol. 3766, pp. 211-220, Springer-Verlag GmbH, 2005.

[4] Yiğithan Dedeoğlu, B. Uğur Töreyin, Uğur Güdükbay, A. Enis Çetin, Silhouette-based Method for Object Classification and Human Action Recognition in Video, *Int. Workshop on Human-Computer Interaction*, Graz, Austria, (in conjunction with ECCV 2006). *Lecture Notes in Computer Science*, vol. 3979, pp. 64-77, Springer-Verlag GmbH, 2006.

[5] C. Canton-Ferrer, J. R. Casas, M. Pardàs. Human Model and Motion Based 3D Action Recognition in Multiple View Scenarios. European Signal Processing Conference (EUSIPCO). Firenze (Italy), September 4-8, 2006.

[6] A. F. Bobick and J. W. Davis, The Recognition of Human Movement Using Temporal Templates, IEEE Trans. on Pattern Anal. and Machine, vol. 23, pp. 257-267, Mar. 1999.

[7] J. Han and B. Bhanu, Individual Recognition Using Gait Energy Image, IEEE Trans. on Pattern Anal. And Machine, vol. 28:2, pp. 316{322, Feb. 2006.

[8] M. K. Hu, Visual Pattern Recognition by Moment Invariants, IRE Trans. on Information Theory, vol. 8:2, pp. 179{187, Feb 1962.

[9] T. Xiang and S. Gong, Beyond Tracking: Modelling Activity and Understanding Behaviour, Int. Journal of Computer Vision, vol. 67:1, pp. 21-51, Feb. 2006.Bowman,

[10] C. Beleznai, B. Frühstück and H. Bischof, *Human Detection in Groups Using a Fast Mean Shift Procedure*, IEEE Int. Conf. On Image Processing, Singapore, Oct. 24-27, 2004

[11] C. Beleznai, B. Frühstück and H. Bischof, *Model-based Occlusion Handling for Tracking in Crowded Scenes*, Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition, 11-13 May 2005, Veszprém, Hungary

[12] C. Beleznai, B. Frühstück and H. Bischof, *Tracking Multiple Humans using Fast Mean Shift Mode Seeking*, Workshop on Performance Evaluation of Tracking Systems - PETS 2005, Breckenridge, Colorado, USA, January 7, 2005

[13] C. Beleznai, B. Frühstück and H. Bischof, *Human Tracking by Fast Mean Shift Mode Seeking*, Journal of MultiMedia, Vol. 1, Issue 1, April 2006, pp. 1-8

[14] T. Schlögl, C. Beleznai, M. Winter and H. Bischof, Performance Evaluation Metrics for Motion Detection and Tracking, Int. Conf. On Pattern Recognition 2004, Cambridge, UK, 23-26 August 2004

[15] C. Beleznai, T. Schlögl, H. Ramoser, M. Winter, H. Bischof and W. Kropatsch, *Quantitative Evaluation of Motion Detection Algorithms for Surveillance Applications*, AAPR/ÖAGM Workshop, Laxenburg, Austria, June 2003

[16] C. Stauffer, W. Eric L. Grimson: Learning patterns of activity using real-time tracking, IEEE Trans. PAMI, Vol. 22(8), pp. 747-757, 2000

[17] E. L. Andrade, R. B. Fisher, S. Blunsden: Detection of emergency events in crowded scenes, Proc. of IEE Int. Symp. on Imaging for Crime Detection and Prevention, pp. 528-533, 2006.

[18] W. Reichardt, Autocorrelation, a principle for the evaluation of sensory information by the central nervous system, pp. 303–317, MIT Press and John Wieley & Sons., New York, NY, 1961.

[19] C. Higgins, Analog VLSI Implementation of Spatio-Temporal Frequency Tuned Visual Motion Algorithms, IEEE Transactions on Circuits and Systems I, vol 52, no. 3, pp. 489-502, March 2005