Autonomous Intelligent Surveillance Networks

Richard Comley

Middlesex University

London, NW4 4BT

r.comley@mdx.ac.uk

***Abstract -* The availability of cheap cameras capable of providing high quality images has prompted their widespread use in video surveillance applications. The need to process the vast number of images produced has created a requirement for intelligent image processing systems capable of largely autonomous operation. The initial identification and subsequent tracking of individuals in busy environments represent two of the most demanding imaging processing problems, for such systems. A review of the major technologies and techniques is presented in this paper, which concludes with a proposal for a network based on an array of intelligent cameras that have the ability to configure themselves into dynamic clusters. This has the potential to enhance significantly the reliability of the tracking algorithms, especially under such circumstances as when discontinuities occur between adjacent views.**

I. INTRODUCTION

The prospect of fully automated visual recognition systems, that can mimic and even surpass the human visual system, has been the long term goal for many researchers, e.g. industrial inspection, medical image processing, surveillance and robotics. However, despite major advances in camera technology, image processing techniques and algorithms and the availability of high speed processors, it remains as elusive today as the paperless office!

Many of the problems are generic but there are also specific, application dependent challenges. Arguably, one of the most complex automated visual recognition problems is that of tracking individuals in crowded environments, e.g. a busy shopping mall. The ‘random’ motion, frequent occlusion, changes in ambient lighting, variations in posture and real-time demands all conspire to compromise the performance of even the most complex systems. Such simple factors as the location of the camera(s) can have a very significant impact on the performance of such systems. For example, an overhead camera is less likely to suffer from problems of occlusion but has only limited features to work with and

only a very limited field of useful operation. Cameras mounted at lower levels generally have a bigger object to work with but will suffer from greater occlusions and variations in light intensity. The normal solution to this problem is to employ multiple cameras and then to correlate between images.

The use of multiple cameras, while providing a richer data source on which to base the visual tracking algorithms, poses further problems of how and where the processing is to be performed, i.e. centralised or distributed. This in turn raises issues for the communication network in terms of its architecture, operation and bandwidth.

Surveillance has two fundamental requirements, initial identification and subsequent tracking of individuals (targets). Identification generally requires access to high quality images from which features can be extracted and used for comparison with a stored dataset. Tracking algorithms, or visual motion detection (VMD) algorithms, can function with lower quality images but require a sufficiently fast frame rate and speed of execution, to produce a reliable output.

Current techniques and technologies suited to video surveillance application are reviewed in this paper. From this, a system based on a distributed array of intelligent cameras is proposed that simultaneously employs a combination of both infrastructure and ad hoc wireless networks. It differs from other systems in that, while it is based on a cluster, or cellular structure, the cells are dynamic in shape and size.

II. TARGET IDENTIFICATION

The identification of features from images has been the subject of intensive research over many years. Some of the most advanced techniques have been developed for use in medical image processing, as an aid to diagnosis. For surveillance purposes, facial recognition is of paramount interest, mimicking human behaviour. Numerous techniques have been developed, the most reliable being based on the detection and quantification of facial ‘landmarks’ (e.g. shape, distance between eyes, etc.).

There are two predominant methods in common use, principal component and linear discrimination analyses. Principal Component Analysis, or feature based methods, were the earliest techniques, often known as eigenfaces, and were pioneered by M.A. Turk and A.P. Pentland [1]. This method identifies key features of an image and disregards all other information. A good example of this technique is described by O. Arandjelo-vic and A. Zisserman [2], which details robust methods to deal with the three major problems; isolation of the face from the background, realignment to deal with differences in orientation, and compensation for variation in expression and partial occlusions. The search space was limited by confining recognition to the characters in a typical movie, but very high success rates (in excess of 90%) were reported.

Linear Discrimination Analysis is an alternative class of algorithms, that use statistical methods to classify features [3]. Various features are categorised into different classes and comparison is performed on the basis of maximising the differences between classes while minimising those of objects within the same class.

A more recent technique, known as Elastic Bunch Graph Mapping [4] uses a Gabor wavelet transform that projects the main features of a face onto a dynamic link architecture, or elastic grid. Classification is based on the displacements of the nodes that make up the grid. Very good results are reported for this method, but it requires accurate initial feature (landmark) information in order to achieve reliable recognition. This normally requires that a number of views of the face to be available, from which to define the landmarks and so is not ideally suited for real-time surveillance.

Having analysed and quantified suitable features, it is then necessary to compare these with data held in a database, in order to complete the recognition process. This represents a non-trivial task, especially if the database is extensive and the comparison has to be achieved in real-time, as generally demanded for surveillance purposes. There is, as yet, no agreed standard on how images are stored for subsequent searching and identification. This complicates the problem enormously.

One very promising technique is described by K. J. Oiu, J. Jiang, G. Xiao and S.Y. Irianto [5]. It is designed to operate directly on JPEG compressed images, without the need for initial decompression. This offers a significant speed advantage, especially when large databases need to be searched. Three block-edge patterns are extracted from the DCT domain and used as indexing keys on which to perform subsequent searches. It has been designed and tested as a means to discern between different types of image content (e.g. car vs. person) and has produced good results, outperforming alternative methods. There is no data on how well it performs in terms of its ability to discriminate between similar objects, e.g. human faces.

An extensive survey of face recognition techniques is provide by W. Zhao, R. Chellappa, P. Phillips and A. Rosenfeld [6].

The above discussion shows that face recognition requires access to high quality images and pre-supposes that a match exists on a database. There are many instances in surveillance applications where neither will be the case. To account for these situations, other features need to be considered, e.g. silhouette, height, colour, texture and motion. Low resolution images have proved to be sufficient for reliable tracking of targets defined in this way [7][8]. Of these, motion is by far the most important.

III. VISUAL MOTION DETECTION

Motion detection is at the heart of most surveillance systems. It has an important role in the initial identification of targets and is clearly essential for any tracking operation.

A major class of Visual Motion Detection (VMD) algorithms are based on a subtraction technique, whereby the background is assumed to be fixed, and unchanging (a reasonable assumption for a fixed camera) and the objects to be tracked move in this fixed space, i.e. temporal information is extracted through the subtraction of successive spatial frames. One of the earliest, and most successful techniques is that of edge detection, with the Canny method [9] being one of the most well know. Edge detection is of great importance as it significantly reduces the amount of data by removing redundant information in the image, while preserving the important structural properties. This is of great importance for any future processing of the image, as it can lead to a huge reduction the computational complexity for subsequent operations. Numerous publications have appeared since this early work, e.g. [10][11], offering various enhancements.

The edge detection method is both computationally efficient and reliable but, as stated above, it assumes a static background. This is not generally the case when crowded scenes are to be analysed. This introduces a significant complication when attempting to track a single object. Optical Flow methods have been used with some success when dealing with moving backgrounds, but these are complex and computationally demanding and, in general, not suitable for real-time implementations.

The most successful models in current use attempt to characterise human motion through a decomposition of the human body into a number of moving regions, or parts. Each of these regions is then assigned a label which allows each to be tracked individually in time. The work of Y. Songy, L. Goncalvesy and P. Perona [12] reports such a system in which the energy patterns of each of the labelled regions is used to classify different actions and activities. It is restricted to triangulated models and is reported to yield a fast and robust method that copes well with extraneous clutter and undetected (occluded) body parts.

The work of G. Brostow and R. Cipolla [13] on a Bayesian detection model, applies the decomposition method to the detection and movements of individuals in crowds. In this, the shapes are decomposed (c.f. [12]) and labeled and the various elements are associated into clusters. Each of these clusters corresponds to the decomposed elements that together make up an individual person. This has been reported to produce highly reliable results when used to detect and track motion, but is does not match the performance of the more specific tracking method reported by T. Zhao and R. Nevatia [14][15] in terms of its ability to discern between individuals entities.

Once we have reliable motion data, the reliability and autonomous operation of the system can be enhanced through the addition of models of behaviour patterns. Certain patterns can be classified as ‘suspicious’, e.g. running, loitering, repeated sightings, association between targets, etc. For example, London Underground has been running trials of an Intelligent Pedestrian Surveillance system, for a number of years [16].

IV. AUTONOMOUS OPERATION

In order for a system to operate autonomously, a degree of intelligence is required. This again has been a classical dilemma for computer scientists since the earliest of times. The origins of AI can be traced back to the pioneering work of Alan Turin. The dilemma for any automated image recognition algorithm is the need for precisely defined models in order that accurate classification can be achieved (e.g. to distinguish between a cat and a dog) yet with sufficient flexibility to deal with changes in orientation, posture, lighting, etc. Numerous AI techniques have been applied to the analysis of video data, the most successful being based on stochastic models. A stochastic process may be defined as one that estimates the likelihood of possible outcomes by allowing for random variations in its inputs over time. This makes it a good candidate for visual recognition problems. Indeed, there is good evidence to suggest that the human visual system operates on this basis and there is much reported research that builds on this presumption, e.g. [17]. The most popular and successful stochastic models in current use are generally based on Bayesian networks and Markov chains.

A Markov model is a probabilistic technique for learning and matching activity patterns. They may be considered automata for which each new state is determined by the prior probability of a process starting from that state and a set of transition probabilities describing the likelihood of a process moving into a new state. They are hence ideally suited to the detection of ‘unusual’ changes in a scene. A Bayesian network, by contrast, is a probabilistic graphical model that represents a set of variables and their probabilistic interdependencies. They are hence suited to applications such as VMD (see [13] earlier). Both models are clearly suited to different aspects of automated surveillance systems.

P. Remagnino, A. Shihab and G. Jones [18] report a highly scaleable, multi-agent distributed architecture, based on stochastic models, suitable for use in surveillance applications. The paper describes the use of modular software (agents) that is able to build a model of a scene by merging information gathered over a period of time (i.e. the training data) which is then used to identify transient objects within the scene. Good success is reported when tested in a car park environment.

An important feature of the systems reported in [18] is its scalability. This is an important property for video surveillance networks, which need to be both flexible and easily expandable, in order to cope with changing demands.

V. CAMERA ARRAYS

Almost all surveillance systems employ multiple camera architectures. These may range from a small number of units that cover key points (e.g. entrance, corridors, etc.) up to large arrays, involving hundreds of units. For tracking purposes, we are invariably dealing with the latter. The next challenge is therefore how to correlate the information obtained by individual cameras, in order to track targets as they move around in the surveillance space. This task is further complicated by such factors as different cameras supplying images in different resolutions and frequently in different formats [19][20].

The kinetics of a target, extracted as part of the VMD process, can provide valuable data, such as bearing and velocity, enabling predictive capabilities to be incorporated, to enhance reliability as the target moves between cameras.

The image processing and correlation can be performed centrally or on a distributed basis, the challenge is to determine which is most appropriate. This is a classical computing science problem with which computer scientists have been grappling for decades. The classical centralised surveillance system is typified by large, central control rooms, having numerous TV screens with a number of human observers. The advantage of such systems is that all the data is available at one location, where extensive computing resources may be deployed. The major disadvantage is the demands placed on the associated communication network and concerns over a single point of failure. A distributed system, by contrast, has only limited computational power at each of the sites, but has a much reduced communication requirement and an inherent robustness to the failure of individual elements. The demands and often localised nature of surveillance tracking problems make a distributed architecture the best choice for such applications. However, initial facial recognition is likely to be better performed on a centralised system drawing on a single image database for searches. The ideal architecture would hence appear to be a hybrid of the two, arranged on a client-server basis, with the distributed elements having a high degree of autonomy, only requesting the assistance of the centre on an infrequent basis.

Advances in sensor area networks and mobile communications networks can contribute much useful knowledge on which to construct the next generation of surveillance systems. As stated above, these should be highly distributed, intelligent networks of cameras and processing units. The work of R. Aguilar-Ponce, A. Kumar, J. Tecpanecat-Xihuit and M. Bayoumi [21] describes such a network based on a hierarchical architecture, comprising a number of clusters of wireless cameras, which communicate with Object Processing Units (OPUs). These in turn form clusters that communicate with Scene Processing Units (SPUs). Object detection and tracking is performed by Region Agents (RAs), which are responsible for the creation of Object Agents that define various parameters associated with targets, that are then downloaded to the SPUs for execution. The tracking algorithm is based on a decomposition model, as previously discussed.

VI. ACTIVE CELL ARCHITECTURE

The model proposed here has much in common with [21], described above, however the cameras are now intelligent and configure themselves into ad hoc wireless networks in order to track individual targets. The clusters are dynamic with individual cameras joining and leaving clusters as targets move in and out of their field of view. The mobile clusters (or cells) are supported by a fixed network of Surveillance Command Modules (SCMs) which communicate with the Intelligent Camera Units (ICUs) and other SCMs. A centralised Master Control Centre (MCC) oversees the whole network and provides high level services. The fixed element of the network is directly analogous to the mobile telephone network and, indeed, it is possible that all intercommunication at this level, may take place via the installed mobile infrastructure. The overall concept is to enable highly flexible and scalable networks of cheap ICUs to be quickly created and rearranged in response to transient demands (e.g. parades, sporting events, etc.).

A major requirement for such is network is how the ICUs will collaborate in this ad hoc manner. When an ICU first joins a network, it will begin to broadcast status information, as any wireless unit joining an ad hoc wireless network [22]. It will also begin the process of analysing the images that it captures, classifying the various objects passing through its view. As objects enter or leave its field of view, key parameters will be communicated to other ICUs, within range. The other ICUs will be doing likewise and it is hence possible through the correlation of this data to establish its identity and orientation with respect to its immediate neighbours. Where an object appears in the field of view of an ICU, it will compare its classification with that of other units within its wireless range, to look for a match. When a match is found, the ICU will know the relative location of the neighbouring ICU. Armed with this spatial awareness, it will be possible for the ICUs to perform the mobile telephone equivalent of a soft handover, whereby adjacent cameras are prepared to receive targets before they appear in their field of view. This has the potential to enhance significantly the reliability of the tracking algorithms, especially in circumstances such as discontinuities between adjacent views (i.e. target disappears from one view before it appears in another).

The spatial data collected by the ICUs will be monitored by the SCMs which will allow a global set of spatial co-ordinates to be established, for use at the higher levels. This is directly analogous to the routing tables created by modern network routing and switching devices. There is hence great scope to draw on research into routing algorithms to support and inform the proposed network.

VII. CONCLUSIONS

Significant advances have been made in recent years in all aspects of the technologies required to create truly intelligent surveillance networks, capable of largely autonomous operation. In order to realise such systems requires the integration of these technologies, which still poses significant problems. The major technological issues have been considered in this paper, but a vast quantity of specialist research exists which it has not been possible to include here. There remains much work to be done before such systems become a reality.

In particular, work is required to establish standards appropriate to the needs of the surveillance community. This is not just the image standards referred to in this paper. There is also an urgent need for debate on the ethical and legal implications for the creation and deployment of such intelligent networks and the development of suitable safeguards and legislation to protect the rights of ordinary citizens. Without this, the deployment of such networks will be viewed with suspicion, which will hamper their effectiveness in the service of society at large.

REFERENCES

[1] M.A. Turk and A.P. Pentland, “Face Recognition Using Eigenfaces”, *Proc IEEE*, pp587-591, 1991.

[2] O. Arandjelovic and A. Zisserman, “Automatic Face Recognition for Film Character Retrieval in Feature-Length Films”, *Proc. of Computer Vision and Pattern Recognition, 2005. CVPR 2005*. IEEE Computer Society Conference, Vol 1, pp860-867, June 2005.

[3] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, “**Regularized discriminant analysis for the small** sample size problem in face recognition”, *Pattern Recognition Letters*, Vol 24, No 16, pp3079-3087, 2003.

[4] J. R. Beveridge, D. Bolme, B. A. Draper and M. Teixeira, “The CSU Face Identification Evaluation System - Its purpose, features, and structure”, *Machine Vision and Applications*, Vol 16, No 2, Springer-Verlag, pp128-138, February 2005

[5] K.J. Qiu1, J. Jiang, G. Xiao, and S.Y. Irianto, “DCT-Domain Image Retrieval Via Block-Edge-Patterns”, *Proc. of International Conference on Image Analysis and Recognition, ICIAR 2006*, Springer-Verlag,

LNCS 4141, pp673-684, 2006.

[6] W. Zhao, R. Chellappa, P. Phillips and A. Rosenfeld, Face recognition: A literature survey, *ACM Computing Surveys (CSUR)*, Vol 35 Issue 4, December 2003.

[7] R. Muñiz and J.A. Corrales, “Novel Techniques for Color Texture Classification”, *Proc. of the 2006 International Conference on Image Processing, Computer Vision and Pattern Recognition*, Vol 1, pp114-120, CSREA Press, 2006,

[8] J. Annesley, V. Leung, A. Colombo, J. Orwell and S. Velastin, “Fusion of Multiple Features for Identity Estimation”, *Proc. of the IET Conference on Crime and Security*, pp534-539, June 2006.

[9] J. Canny, “A computational approach to edge detection”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol 8 No 6, pp679-698, 1986.

[10] J. Oh and C. Jeong, “The DSC Algorithm for Edge Detection”, *Proc. of the 17th Australian Conference on Artificial Intelligence, AI 2004*, Vol 3339/2004, Springer-Verlag, pp 967-972, December 2004.

[11] D. Knossow, J. van de Weijer, R. Horaud, and R. Ronfard, “Articulated-body Tracking through Anisotropic Edge Detection”, *Lecture Notes in Computer Science*, Volume 4358/2007, Springer-Verlag, pp86-99, May 2007

[12] Y. Songy, L. Goncalvesy, and P. Perona, “Learning Probabilistic Structure for Human Motion Detection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 25, Issue 7, pp 814-827, July 2003.
[13] G. Brostow and R. Cipolla, “Unsupervised Bayesian Detection of Independent Motion in Crowds”, *Proc. of* *Computer Vision and Pattern Recognition, IEEE Computer Society Conference*, pp 594-601, June 2006.

[14] T. Zhao and R. Nevatia. “Tracking multiple humans in complex situations”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 26 No 9, pp1208-1221, 2004.

[15] T. Zhao and R. Nevatia. “Tracking multiple humans in crowded environment”. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2004),* Vol 2 Issue 27, pp406-413, July 2004.

## [16] J. Hogan, “Smart software linked to CCTV can spot dubious behaviour”, *New Scientist*, July 2003.

[17] R. Rao and D. Ballard, A Class of Stochastic Models for Invariant Recognition, Motion, and Stereo, *Technical Report: NRL96.1*, University of Rochester, 1996

[18] P. Remagnino, A. Shihab and G. Jones, Distributed intelligence for multi-camera visual surveillance, Pattern Recogition, Elsevier Science, Vol 37 Issue 4, pp675-689, April 2004

[19] J. Kang, I. Cohen, and G. Medioni, “Multi-views tracking within and across uncalibrated camera streams”, *Proc. First ACM SIGMM International Workshop on Video* *Surveillance*, (New York), pp. 21-33, ACM Press, 2003.

[20] T. Gandhi and M. M. Trivedi, “Calibration of a reconfigurable array of omnidirectional cameras using a moving person”, *Proc. 2nd ACM International Workshop on Video* *Surveillance and Sensor Networks*, (New York), pp12-19, ACM Press, 2004.

[21] R. Aguilar-Ponce, A. Kumar, J. Tecpanecat-Xihuit and M. Bayoumi, “A network of sensor-based framework for automated visual surveillance”, *Journal of Network and Computer Applications,* Vol 30,  Issue 3*,* pp 1244-1271, August 2006.

[22] C. Murthy, B. Manoj, Ad Hoc Wireless Networks - Architectures and Protocols, Pearson Books, June 2004. ISBN10: 013147023X