An automatic drowning detection surveillance system for challenging outdoor pool environments

How-Lung Eng, Kar-Ann Toh, Alvin H. Kam, Junxian Wang and Wei-Yun Yau
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore
hleng@i2r.a-star.edu.sg

Abstract

Automatically understanding events happening at a site is the ultimate goal of visual surveillance system. This paper investigates the challenges faced by automated surveillance systems operating in hostile conditions and demonstrates the developed algorithms via a system that detects water crises within highly dynamic aquatic environments. An efficient segmentation algorithm based on robust block-based background modeling and thresholding-with-hysteresis methodology enables swimmers to be reliably detected amid reflections, ripples, splashes and rapid lighting changes. Partial occlusions are resolved using a Markov Random Field framework that enhances the tracking capability of the system. Visual indicators of water crises are identified based on professional knowledge of water crises detection, based on which a set of swimmer descriptors has been defined. Through seamlessly fusing the extracted swimmer descriptors based on a novel functional link network, the system achieves promising results for water crises detection. The developed algorithms have been incorporated into a live system with robust performance for different hostile environments faced by an outdoor swimming pool.

1. Introduction

Automated video-based surveillance for real-time human behavior analysis provides an efficient way of detecting the occurrence of any abnormal events amid our surroundings. The technical challenges faced encompass the need to reliably detect and track moving targets within a possibly dynamic background and an inference module that interprets targets' behavioral patterns as events with semantic meaning. The recent increase in demand for such technology in real applications, e.g., for homeland security, motivates research into systems that extend beyond indoor and controlled outdoor environments [1]-[3] to more realistic hostile environments typically encountered in real applications [4],[5]. Major difficulties however continued to be faced by most state-of-the-art systems due to rapidly changing environmental lighting, highly dynamic backgrounds and poor visibility of targets.

In this paper, an outdoor surveillance problem, which involves human behavior monitoring within hostile aquatic environment, is considered. On top of some new insights into problems faced for common outdoor environments, problems unique to human detection within dynamic aquatic environment are also detailed. Such system is highly useful for lifeguard and potentially be applied at unattended pool to enhance the safety at swimming pool.

Previous work on automated aquatic surveillance system for drowning detection is limited to a few patented systems given in [6]-[8]. The reliance on underwater cameras in these systems inherit weaknesses: 1) expensive installation costs, and 2) drowning detection being constrained to victims who have sunk to the bottom of the pool. To circumvent these drawbacks, the proposed system is based on a network of highly mounted overhead cameras. This allows the detection of early drowning behavior from the onset of water crisis situation. Hence, any rescue effort could be initiated much earlier than those in [6]-[8].

In the considered problem, one major technical challenge faced is to accurately detect and track swimmers within the noisy outdoor aquatic environments. The conventional methods [1],[2], where single or mixture of Gaussian distribution was used to model the temporal variation of background pixels, have been found to be inadequate for this highly non-stationary environments. To effectively detect and segment swimmers, a novel block-based background modeling and a thresholding-with-hysteresis methods are developed. The block-based background modeling captures well the spatial dependencies and dynamic nature of the aquatic environment. Whereas, the thresholding-with-hysteresis method addresses the problem of selecting thresholds within a background subtraction framework, which is to yield a high sensitivity in detecting swimmers while suppressing the background noise.

The ability to handle partial occlusion is also incorporated into the system with the development of a novel occlusion handling scheme. In contrast to previous work based

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

on spatial Mahalanobis distance [9] and geometrical features [10], the proposed method captures the spatial and temporal correlation of swimmers in addition to color information based on a Markov Random Field (MRF) framework to yield better performance. Promising results for water crises detection have been achieved using a unique functional link network which fuses extracted swimmer descriptors in an optimal way. This has been proven to be a superior descriptor fusion method compared to the hierarchical method proposed in [15],[16].

This paper is organized as follows: Section 2 describes technical challenges faced in human detection within aquatic environments. Section 3 details the proposed block-based background model and thresholding-with-hysteresis method for swimmer extraction. Section 4 explains the proposed Markov Random Field (MRF) framework for partial occlusion handling. Section 5 describes the proposed functional link network for water crises inference. Experimental results and some concluding remarks are presented in Sections 6 and 7, respectively.

2 The challenges at an outdoor aquatic environment

Due to continual disturbances caused by water ripples and splashes, the aquatic environment is relatively more hostile than most indoor and outdoor environments typically considered in the literature. Figure 1 shows two consecutive frames taken from a typical pool scene. The background movements at the reflective regions, lane dividers and shadows could be easily mis-identified as foreground objects' movements. In addition, poor visibility of swimmers in water due to reflections (from sunlight and nighttime lighting) and the problem of occlusion makes accurate segmentation a very challenging problem.

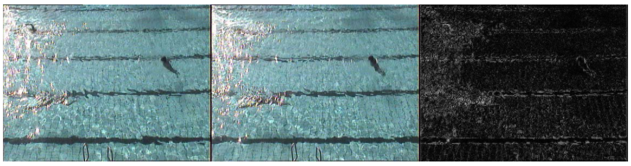


Figure 1: Background movements at reflective areas, lane dividers and shadows. From left to right: Two consecutive frames from a typical scene captured and the corresponding absolute difference image between both frames.

Apart from the above issues unique to aquatic environments, there are common problems faced in outdoor surveillance, i.e., continual illumination changes due to ambient lighting, auto-gain effects of the cameras, and etc. Fast background updating is important to adapt to such illumination changes. This however will exacerbate the problem of foreground objects blending into the background model due to segmentation errors. Corruption of the background model will in turn lead to more segmentation errors on the subsequent frames.

Other technical challenges include the need to have an algorithm that runs real-time and is suitable for implementation at low cost, low power and using common-off-the-shelf hardware platforms.

3 Swimmer detection

The proposed methodology for swimmer extraction consists of a learning phase which builds the initial background model and a detection phase which segments out swimmers from the non-stationary pool background as presented in Figure 2.

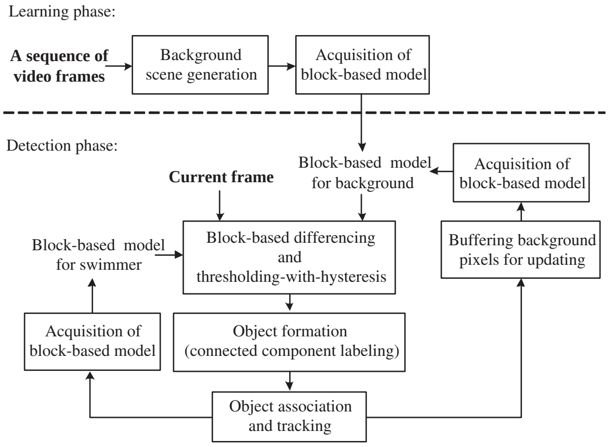

Learning phase:
A sequence of video frames
Background scene generation
Acquisition of block-based model
Detection phase:
Current frame
Block-based model for background
Acquisition of block-based model
Block-based model for swimmer
Block-based differencing and thresholding-with-hysteresis
Buffering background pixels for updating
Acquisition of block-based model
Object formation (connected component labeling)
Object association and tracking

Figure 2: Architecture of the proposed segmentation algorithm.

3.1 Initial block-based background modeling

Video frames captured during the learning phase possibly contain both moving and stationary foreground swimmers. Removing foreground swimmers to establish a "clean" empty scene is crucially needed to yield an accurate background model for good segmentation of swimmers. To build a clean background scene, a skin color model is first applied as a pre-processing step to isolate swimmer pixels from the background formation process. Residue swimmer pixels, which could be regarded as impulsive noise among the collected background pixels when analyzed over time, are then effectively removed using a temporal vector median filter as follows.

Let $\mathbf{V}_t(i,j)$ be an array of color vectors collected over T frames, i.e. $\mathbf{V}(i,j) = \{\mathbf{V}_t(i,j) \mid t = 1,\ldots,T\}$, where $\mathbf{V}_t(i,j)$ is the color vector of the tth image at position-(i,j). The sampling rate that determines the temporal interval between two consecutive $\mathbf{V}_t(i,j)$ was decided empirically after considering a tradeoff between the duration needed for the learning phase and the efficiency to remove

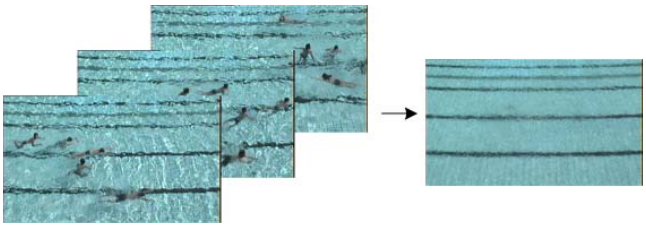


Figure 3: Generation of a background scene using a temporal vector median filter. (a) A sequence of frames containing foreground swimmers; (b) Background scene obtained.

swimmer pixels. Performing vector median filtering on $\mathbf{V}_t(i,j)$ for all i and j produces $\mathbf{Y}(i,j)$ such that

$$\mathbf{Y}(i,j) = \Big\{ \mathbf{V}_q(i,j) \in \mathbf{V}(i,j) | \min_{\mathbf{V}_q(i,j)} \sum_{p=1}^{T} |\mathbf{V}_p(i,j) - \mathbf{V}_q(i,j)| \Big\}. \tag{1}$$

A composition of $\mathbf{Y}(i,j)$ forms the initial background scene

$$\mathbf{B} = \Big\{ \mathbf{Y}(i,j) \mid i = 1, \ldots, N \text{ and } j = 1, \ldots, M \Big\}, \tag{2}$$

where N and M are the height and width of the video frame respectively. Figure 3 shows the generation of a “clean” background scene from a sequence of “training” frames containing moving and stationary swimmers.

Instead of modeling each background pixel independently, the background model is defined to be cluster centroids of homogeneous color regions within blocks forming the background scene $\mathbf{B}$. This captures the strong spatial correlation among the pixels. From our experiments, we found CIELa*b* to produce better swimmer segmentation results compared to other color spaces. The background scene is thus first converted into the CIELa*b* space, forming matrix $\mathbf{B}'$. To form the background model, two steps are then taken as follows: (i) dividing $\mathbf{B}'$ into $m \times n$ number of non-overlapping $s \times s$ square blocks, and (ii) applying a hierarchical k-means [12] on each square block to obtain cluster centroids of homogeneous regions. The clustering process is initiated by assuming each square block to be one dominant data cluster. In the subsequent iterations, smaller and more compact data clusters are formed through splitting until the distance of the two closest cluster centers is smaller than a threshold or the number of clusters reaches three, whichever is achieved first. The initial background model is thus defined to be

$$\mathbf{C} = \big\{ \mathbf{C}(i,j) \mid i = 1, \ldots, m;\ j = 1, \ldots, n \big\}, \tag{3}$$

where $\mathbf{C}(i,j) = \{\mathbf{C}_k(i,j)\}$ is the set of cluster centroids of homogeneous regions formed in block-(i,j).

3.2 Foreground object detection

Swimmers are detected within a modified background subtraction framework. Color discrepancy is defined to be the minimum l_1-norm between the pixel of the incoming image and cluster centroids $\mathbf{C}(i,j)$ within the corresponding background block. Additionally, distances from the surrounding eight-connected background blocks are also computed to account for background motions.

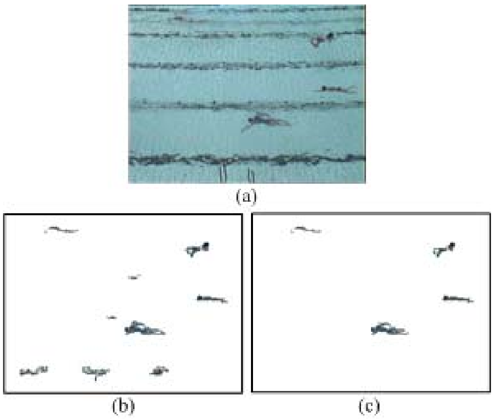


Figure 4: Comparison of segmentation results between cases (b) without and (c) with the eight-neighboring background blocks when computing $\mathbf{D}$ for the sample frame in (a).

Let $d_{min}(i,j)$ be the minimum color discrepancy measure between a pixel color at location-(i,j) and the cluster centroids of its surrounding eight-connected background blocks. Obtaining $d_{min}(i,j)$ for all i and j yields a difference image $\mathbf{D}$ to be thresholded for swimmer detection:

$$\mathbf{D} = \Big\{ d_{\text{min}}(i,j) \mid i = 1, \ldots, M\ ;\ j = 1, \ldots, N \Big\}. \tag{4}$$

With the consideration of the eight-neighborhood structure, this reduces significant mis-classification errors at lane dividers and shadows due to the background movements. Figure 4 depicts a typical example obtained for cases with and without the eight-neighborhood structure.

The choice of a suitable threshold has always been a tricky problem for foreground detection. Single-thresholding method commonly applied for deciding what a significant change is often leads to a tradeoff between the detection rate and mis-classification rate. An elegant solution based on the thresholding-with-hysteresis principle [11] is proposed where a high and a low threshold, denoted T_h and T_l respectively, are used. A swimmer is detected only if a region of connected pixels above T_l contains also a given fraction of pixels above T_h. We first produce a binary thresholded map marking pixels with $d_{min}(i,j) > T_h$ using “1”s. A low resolution image of the binary map, called the *parent map* is then constructed. Elements in this parent map are labelled “1” if the number of original binary pixels within it with $d_{min}(i,j) > T_h$ is above a pre-determined value. A connected region of pixels with d_{min} above T_l is classified as foreground if its corresponding area in the parent map also forms a connected blob. This hierarchical thresholding approach produces high detection sensitivity while minimizing the mis-classification of small moving background as foreground.

In the same way as forming the background model, once a swimmer is detected, a block-based swimmer model is defined and updated every frame to improve the segmentation accuracy. Let $\mathbf{C}^f(i,j) = \{\mathbf{C}^f_k(i,j)\}$ be the set of cluster centroids of k homogeneous regions formed based on foreground pixels within block-(i,j). The foreground model is defined as

$$\mathbf{C}^f = \big\{\mathbf{C}^f(i,j) \mid i = 1,\ldots,m;\ j = 1,\ldots,n\big\}. \quad (5)$$

The color discrepancy measure defined as the minimum l_1-norm between the pixel color of the incoming image and swimmer cluster centroids $\mathbf{C}^f(i,j)$ within the corresponding and also the surrounding eight connected swimmer blocks are computed for all pixels, giving

$$\mathbf{D}^f = \Big\{d^f_{\text{min}}(i,j) \mid i = 1,\ldots,M\ ;\ j = 1,\ldots,N\Big\}\ . \quad (6)$$

Pixels with $d^f_{\text{min}}(i,j) < d_{\text{min}}(i,j)$ are classified as the highest confidence swimmer pixels. The remaining swimmers pixels are detected using the thresholding-with-hysteresis scheme detailed earlier.

3.3 Background model updating

There is a need to recursively update the background model to adapt to dynamic environmental changes. If the background is not updated fast enough, rapid lighting change for example will cause accumulative errors between updates. These errors adversely affect the background updating process itself, thus generating a negative performance spiral. Thus, this rule out any gradual updating scheme based on "learning" parameters.

For our system, background model (defined in (3)) is updated by operating on the mean of background pixels accumulated for 10 consecutive samples. Although the median is probably a more robust operator compared to the mean, it could not be implemented within the real-time operation constraint imposed. Besides, our segmentation algorithm is sufficiently robust in ensuring the mean operation does not introduce perceptible degradation to the background updating process.

4 Partial occlusion handling

Detected foreground regions are resolved to be individual swimmers using conventional connected component analysis. These swimmers are then tracked based on minimum spatial Mahalanobis distances between consecutive frames. When occlusions occur, the merging of a few swimmers into one single blob introduces challenging problems for the tracking process. We developed a unique MRF framework for resolving partial occlusion while relying on a linear prediction scheme to determine centroids of individual swimmers when full occlusion happens.

The partial occlusion handling algorithm involves two stages: 1) over-segmenting the single detected blob into homogeneous regions based on a hierarchical k-means clustering algorithm [12], and 2) labeling each segmented small regions to the corresponding swimmers involved in the occlusion based on a MRF framework. In oversegmenting the single blob, a hierarchical k-means clustering algorithm groups pixels within the single blob into k different clusters based on pixel intensity. In spatial domain, this forms small homogeneous regions, where each small region has intensity different from its neighboring regions. To eliminate spurious noise, any region with size smaller than T_s (typically, $T_s <$ 20) is merged with other neighboring regions.

Let $\mathbf{X}$ be the set of all possible labels for the homogeneous regions, $\mathbf{X} = \{\mathbf{X}_p = x_p \mid p \in \Omega,\ x_p \in \Phi\}$, where $\Omega = \{1,\ \ldots,K\}$ represents the set of region index and Φ is the set of swimmers involved in the occlusion. Let $\mathbf{Y}$ be the set of all observed gray values of all homogeneous regions. The labeling process in the second stage is to identify an optimal labeling x^* for a given $\mathbf{Y} = y$, by maximizing

$$x^* = \underset{x}{\text{argmax}}[\ln P(\mathbf{Y} = y\,|\mathbf{X} = x) + \ln P(\mathbf{X} = x)]. \quad (7)$$

Assuming the gray level distribution of each region to be independent, the likelihood function $P(\mathbf{Y} \mid \mathbf{X})$ is thus given by

$$P(\mathbf{Y} = y \mid \mathbf{X} = x) = \prod_{p\in\Phi} P(y_p \mid x_p), \quad (8)$$

where the $P(y_p \mid x_p)$ is the conditional probability that the region would be labeled x_p based on an observed gray level distribution y_p. Let $\mathcal{H}$ be the normalized color histogram with respect to the peak in the intensity space. Let also $\mathcal{H}^s$ and $\mathcal{H}^r_p$ be the normalized histogram for swimmer before occlusion and segmented region-p, respectively. The conditional probability that correlates region x_p to the swimmers involved in the partial occlusion is defined as

$$P(y_p \mid x_p) = \Big(\frac{D_p}{\sum_{q\in\Phi} D_q}\Big)^{-1}, \quad (9)$$

where D_p and D_q are the summation of the absolute corresponding bin differences between $\mathcal{H}^s$ and $\mathcal{H}^r_p$, and $\mathcal{H}^s$ and $\mathcal{H}^r_q$, respectively.

To provide spatial and temporal smoothness constraint, $P(\mathbf{X} = x)$ of (7) is given by

$$P(\mathbf{X} = x) = \frac{1}{Z}\text{exp}\Big(\sum_{c\in\mathcal{C}_s} V_{\mathcal{C}_s}(x) + \sum_{c\in\mathcal{C}_t} V_{\mathcal{C}_t}(x)\Big)\ , \quad (10)$$

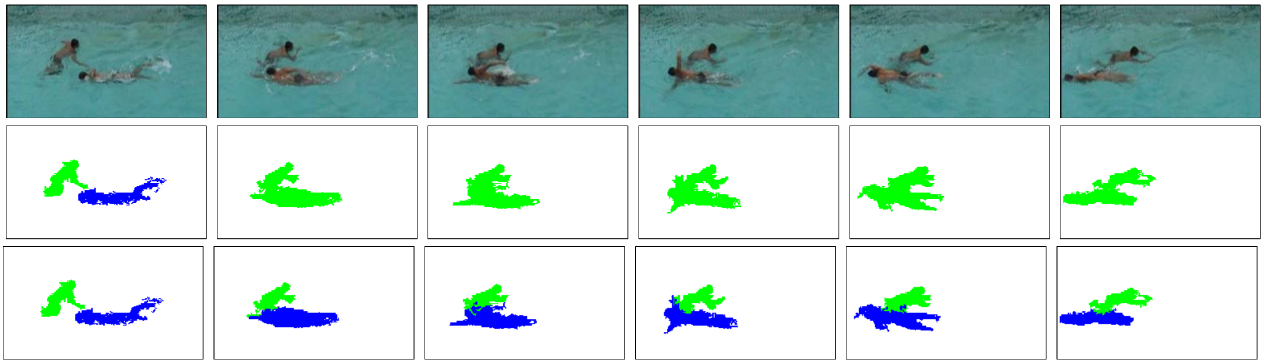


Figure 5: Partial occlusion handling scheme. 1^{st} row: Sample frames of a video sequence containing instances of partial swimmer occlusion; 2^{nd} row: Detected swimmer blobs, which gives only one single blob (with only one colour) when occlusion happens; 3^{rd} row: The proposed scheme resolves the partial occlusion.

where $\mathcal{C}_s(x)$ and $\mathcal{C}_t(x)$ are the set of all possible cliques associated with the neighborhood system, and Z is a normalizing constant. To ensure smoothness in the spatial domain, the spatial clique potential function is defined to be

$$V_{\mathcal{C}_s}(x) = \begin{cases} \beta, & \text{if } x_p = x_q \text{ and } p,q \in \mathcal{C}_s, \\ -\beta, & \text{otherwise.} \end{cases} \quad (11)$$

Temporal smoothness is meanwhile handled by assigning

$$V_{\mathcal{C}_t}(x) = \begin{cases} \alpha \frac{A_u}{\sum_{v \in \Phi} A_v}, & \sum_{v \in \Phi} A_v > 0 \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

where A_u and A_v are the overlapping areas between the region-p and the projected segmentation maps of swimmers u and v from previous frame respectively using a linear prediction scheme. Values of β and α of 0.5 and 5 respectively are found to be robust for almost all partial occlusion cases encountered by our system.

It is computational expensive to search through the complete configuration space of (7) in real time. An iterative deterministic label updating method [13] is adopted to obtain a suboptimal but acceptable solution. Figure 5 shows a typical case of partial occlusion observed at the swimming pool. As shown, the devised scheme is quite effective in resolving these cases with robust results.

5 Water crises inference

Our methodology for water crises inference is to first identify the visual indicators used by the professional lifeguarding community [14]. We then model these established visual indicators by designing special swimmer descriptors that could be extracted from the temporal history of the corresponding swimmer's segmentation map. These descriptors are then optimally fused by using our developed *generalized reduced multivariate polynomials network* (GRM) for water crises inference.

5.1 Swimmer descriptor extraction

A set of five swimmer descriptors is used to model water crises visual indicators. The following provides a summary of swimmer descriptors that has been defined in our previous work [9]:

1. **Speed** (v_i): A swimmer's translational speed is defined as the difference in average centroid positions computed over a one second period. A compensation coefficient is incorporated to account for camera perspective effects.
2. **Posture** (p_i): Posture defines a swimmer's dominant position over a short temporal duration, typically 3 seconds, based on the angle of the principle axis of the best-fit ellipse enclosing the swimmer.
3. **Submersion index** (s_i): A sinking swimmer usually exhibits a higher colour saturation as the reflected light passes through a certain depth of water. Submersion index is defined as difference between a swimmer's current average saturation and the lowest value for it since being tracked.
4. **Activity index** (Ac_i): A swimmer's activity index is defined as the ratio between the cumulative area of pixels covered by the swimmer and the average area of the best-fit ellipse enclosing him over the same duration.
5. **Splash index** (Sp_i): A swimmer's splash index measures the number of splash pixels within a bounding box containing the swimmer.

5.2 Drowning detection using a functional link network

This section details a modified functional link network (FLN) [15],[16] to seamlessly fuse the five swimmer descriptors defined in the previous section for better water crises inference compared to the rule-based technique in our previous work.

Certain FLN algorithms have been known to be a universal approximator and reported to have a faster learning rate than the conventional feedforward/recurrent neural networks [17]. However, the substantial number of parameters to be estimated in the conventional FLN makes the approach less attractive. Instead, a new algorithm, named *generalized reduced multivariate polynomials network* (GRM), is proposed with the number of parameters to be estimated increases *almost linearly* with the model orders and the number of inputs. In contrast to the usual neural network training, our GRM requires only a single training step. The proposed drowning inference involves two stages: 1) an offline stage to obtain an optimal set of weights based on a set of training data, and 2) an on-line stage to incorporate the algorithm into a live system for real time event inference.

Let $\mathcal{S}_{train} = \{\boldsymbol{x}_i \in \Re^p, y_i \in \Re\}$ for $i = 1, ..., m$ be the training set and $\mathcal{S}_{test} = \{\boldsymbol{x}_i \in \Re^p, y_i \in \Re\}$ for $i = 1, ..., n$ be the test set, where $\boldsymbol{x}_i = [v_i, p_i, s_i, Ac_i, Sp_i]^T \in \Re^p$ $(p = 5)$ is the extracted swimmer descriptors and the swimming event is given by y_i. The problem of training is to find the best approximation of y, denoted by $\hat{y}$ which is a function of $\boldsymbol{x}$ and the weights $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_p]$, that minimizes the least square error

$$s(\boldsymbol{x}, y, \boldsymbol{\alpha}) = \sum_{i=1}^{m} e_i^2(\boldsymbol{\alpha}) + b\,\|\boldsymbol{\alpha}\| = \sum_{i=1}^{m} \left(y - \hat{y}(\boldsymbol{x}_i, \alpha)\right)^2 + b\,\|\alpha\|, \tag{13}$$

where $\|\cdot\|$ denotes the l_2-norm and the regularization constant is empirically set to be $b = 10^{-2}$. Minimizing (13) gives the weight parameters to be estimated as follows

$$\boldsymbol{\alpha} = (\hat{\mathbf{X}}^T \hat{\mathbf{X}} + b\mathbf{I})^{(-1)} \hat{\mathbf{X}}^T \mathbf{Y}\ , \tag{14}$$

where $\mathbf{I}$ is a $(p \times p)$ identity matrix, $\mathbf{Y} = [y_1, \ldots, y_m]^T$ and $\hat{\mathbf{X}} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]^T$ for a linear system [18].

Instead of adopting a linear combination, i.e., $\sum_j \alpha_j\, x_{ij}$ given in [18], we consider a r^{th}-order multinomial model which describes better the nonlinear input-output relationship as follows

$$\hat{y}_{MN} = \alpha_0 + \sum_{j=1}^{r} (\alpha_{j1} x_1 + \alpha_{j2} x_2 + \cdots + \alpha_{jp} x_p)^j\ . \tag{15}$$

Through a linearization and generalization process detailed in Appendix A, (15) is extended and generalized including high order terms, forming the GRM model given by

$$\begin{aligned}\text{(GRM):}\quad \hat{y}_{GRM} \;&=\; \alpha_0 + \sum_{j=1}^{r} \alpha_{j+1} (x_1 + x_2 + \cdots + x_p)^j \\ &+ \sum_{w=1}^{r} \sum_{j=w}^{r} (\boldsymbol{\alpha}_j^T \cdot \boldsymbol{x}^w)(x_1 + x_2 + \cdots + x_p)^{j-w}\ ,\end{aligned} \tag{16}$$

where $\boldsymbol{x}^w \triangleq [x_1^w, x_2^w, \cdots, x_p^w]$ and $p\,, r \geq 2$. $\hat{\mathbf{X}}$ in (14) for this model is the Jacobian of $\hat{y}_{GRM}(\boldsymbol{\alpha}, \boldsymbol{x})$.

The GRM classifier is evaluated on the test set $\mathcal{S}_{test}$ which is not used in training.

6 Experimental results

A real-time aquatic surveillance system, consisting of a network of overhead cameras feeding video signals into a computer cluster, has been set up on trial at a public swimming pool for a wide range of testing since past six months. Figure 6 presents a typical scene at the pool with highly dynamic background, making accurate swimmer segmentation, tracking and water crises inference a very challenging task.

6.1 Swimmer detection and tracking

The proposed system runs real-time at 4 frames/second throughout the day and night, under various hostile environments as shown in Figures 6 and 7. Generally, consistently good swimmer detection and tracking results have been obtained, illustrating the robustness of the algorithm operating under numerous real situations amid strong glares, rain and night-time reflections.

Compared to the W4 system [3], our algorithm consistently achieves better segmentation (as shown in Figure 6) in terms of the capability to detect small swimmers and suppressing errors due to the dynamic background. Objectively, a lower segmentation error rate, i.e.:

$$\text{Error rate (ER)} = 100 \times \frac{\text{Error pixel count}}{\text{Frame size}}\ (\%), \tag{17}$$

is also consistently obtained as tabulated in Table 1 for segmentation results shown in Figure 6.

	ER (%)			
Proposed algorithm	2.64	2.10	2.51	1.85
W4 [3]	5.41	3.36	3.63	3.00

Table 1: Error rate for segmentation presented in Figure 6. The ground truth for segmentation was obtained manually.

6.2 Water crisis inference

Three sets of thirty minutes video sequences ($\approx$ 7200 frames) for the respective events, i.e., *drowning*, *normal swimming* and *treading*, were collected with representative examples depicted in Figure 8. From these sequences, a total number of 1000, 1300 and 2000 sets of swimmer descriptors were extracted for the respective drowning, normal swimming and treading events. Realistic water distress was simulated and verified by professional lifeguard. A 10-fold validation process was performed with 90% of each class being selected as $\mathcal{S}_{train}$, while the remaining being assigned as $\mathcal{S}_{test}$, and such process was repeated 10 times with different combinations of $\{\mathcal{S}_{train}\ , \ \mathcal{S}_{test}\}$. The ground truth for the classification was obtained manually by assigning the three events {drowning, treading, normal swimming} to be $\{0, 0.5, 1\}$,

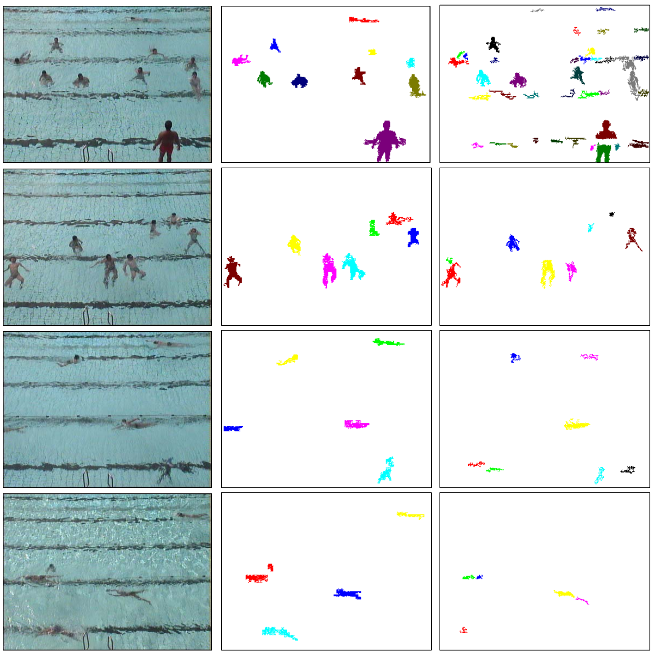


Figure 6: Comparison of segmentation on sample frames captured from different scenarios. 1^{st} column: Samples of scene captured; 2^{nd} column: Segmented swimmers using our proposed algorithm; 3^{rd} column: Segmented swimmers using the well-known W4.

respectively. Figure 9 depicts one of the 10-fold validation in classifying the input data to appropriate classes. Due to significant overlapping of the swimmer attributes between drowning and treading, there are relatively more false classifications between these classes.

As compared to other well known data fusion algorithms, e.g., *optimal weighting method* (OWM) [18] and *feedforward neural network* (FNN) (see e.g., [19]), better classification results have been obtained by our GRM with much smaller error rates for both the training and test data sets as presented in Table 2. GRM has attained an average test error rate of about 5.5% compared with approximate 12% and 13% given by OWM and FNN, respectively. The reason that FNN has large classification error is due to its convergence to local error solution such that the treading case has been mis-classified in some instances. Furthermore, at different threshold selections, our algorithm consistently attains a higher *receiving operating characteristic* (ROC) curve for water crises detection as compared to those of OWM and FNN shown in Figure 10. When plotting the ROC curve, both the normal swimming and treading events were considered as non-drowning case. With about 5% of false acceptance rate, the system achieves at about 90% of correct drowning incident detection.

For practical implementation, it is essential to have a very high authentic acceptance rate so that a genuine water crisis could always be detected. The corresponding increase in false acceptances can always be ignored after visual verification by the professional lifeguard on duty.

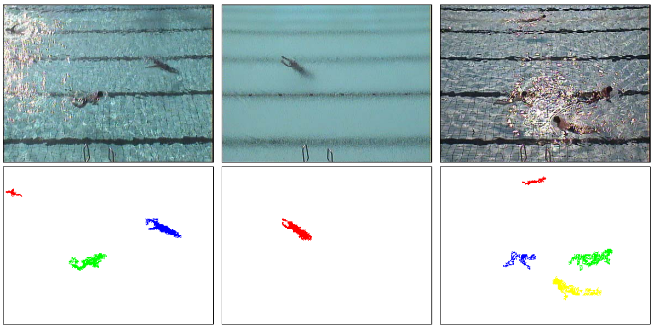


Figure 7: Swimmer detection obtained using our segmentation algorithm at different hostile pool environments: a morning scene with sunlight reflection, a rainy day and a nighttime scene.

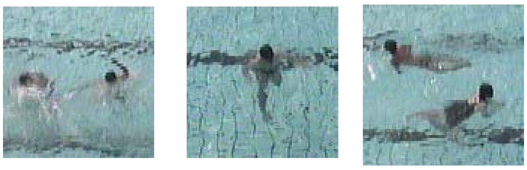


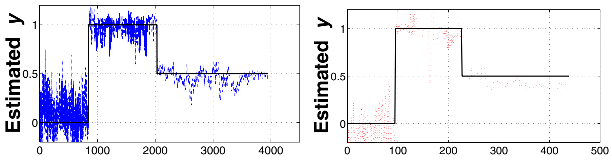
Figure 8: (From left to right) A sample of three different swimming events: *distress*, *treading* and *normal swimming*.

	$\%Error_{train}$ (σ_{train})	$\%Error_{test}$ (σ_{test})
GRM	4.625 (0.131)	5.494 (0.171)
OWM	9.736 (0.167)	11.941 (0.174)
FNN	11.732 (0.161)	12.673 (0.170)

Table 2: Comparison on the average error rate based on a 10-fold validation process.

7 Conclusion and future work

This paper provides insights into automated surveillance within dynamic conditions as demonstrated by a drowning detection system for outdoor public swimming pools. Promising results were shown on all fronts, from novel algorithms that effectively track swimmers amid very hostile conditions, a robust partial occlusion handling scheme to a novel water crises inference module that models professional knowledge of water crises recognition. We believe that our work provides a valuable frame of reference for handling the unresolved technical challenges of practical automated surveillance. Future work for us includes developing a coherent framework and practical algorithms for



(a) Training data (b) Test data

Figure 9: Comparison between the estimated $\hat{y}$ (dashed line) and the truth y (solid line) for training and test data, respectively.

inferring semantic events involving a group of people, for example drowning occurring within a very crowded pool or beach where it is virtually impossible to track everyone individually.

Acknowledgments

This work is supported by the Enterprise Challenge Unit, Singapore's Prime Minister's Office and the Singapore Sports Council.

Appendix A

Given two points $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_1$ on the multinomial function (15) which is differentiable. By the Mean Value Theorem, the multinomial function $f(\boldsymbol{\alpha}) = (\alpha_{j1}x_1 + \alpha_{j2}x_2 + \cdots + \alpha_{jp}x_p)^j$, $j = 2, ..., r$ about the point $\boldsymbol{\alpha}_1$ can be written as:

$$f(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha}_1) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_1)^T \nabla f(\bar{\boldsymbol{\alpha}}), \qquad (18)$$

where $\bar{\boldsymbol{\alpha}} = (1 - \beta)\boldsymbol{\alpha}_1 + \beta\boldsymbol{\alpha}$ for $0 \leq \beta \leq 1$. Let $\boldsymbol{x} = [x_1, ..., x_p]^T$. With appropriate choice of terms based on (18), omitting the coefficients within $f(\boldsymbol{\alpha}_1)$ and $\nabla f(\bar{\boldsymbol{\alpha}})$, and including the summation of weighted input terms gives the following multivariate model:

$$\begin{aligned} \hat{f}_{RM'} &= \alpha_0 + \sum_{j=1}^{p} \alpha_j x_j + \sum_{j=1}^{r} \alpha_{p+j}(x_1 + x_2 + \cdots + x_p)^j \\ &\quad + \sum_{j=2}^{r} (\boldsymbol{\alpha}_j^T \cdot \boldsymbol{x})(x_1 + x_2 + \cdots + x_p)^{j-1}, \qquad (19) \end{aligned}$$

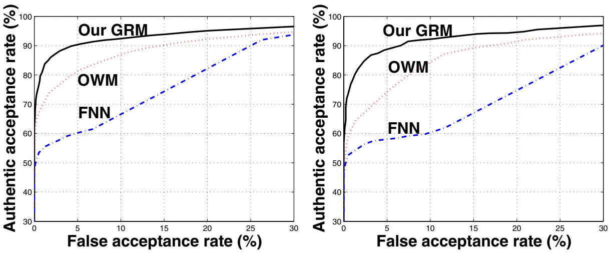
where $p, r \geq 2$. Including more individual high-order terms for (19) yields the following (RM) model

$$\begin{aligned} \hat{f}_{RM} = \alpha_0 + \sum_{k=1}^{r} \sum_{j=1}^{p} \alpha_{kj} x_j^k + \sum_{j=1}^{r} \alpha_{rp+j}(x_1 + x_2 + \cdots + x_p)^j \\ + \sum_{j=2}^{r} (\boldsymbol{\alpha}_j^T \cdot \boldsymbol{x})(x_1 + x_2 + \cdots + x_p)^{j-1}. \qquad (20) \end{aligned}$$

Generalizing (20) by including higher order term of $(\boldsymbol{\alpha}_j^T \cdot \boldsymbol{x})$ up to w^{th}-order gives (16).

References

[1] C. Wren, A. Azarbayehani, T. Darrell, T. and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Machine Intell.*, 19, pp. 780-785, 1997.



(a) Training data (b) Test data

Figure 10: ROC curves for water crises detection obtained using our proposed GRM, OWM [18] and FNN [19] on the training and test data, respectively.

[2] N. Friedman and S. Russell, "Image segmentation in video sequence: A probabilistic approach," *Proc. Thirteenth Conf. Uncertainty in Artificial Intelligence*, pp. 175-181, 1997.

[3] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Machine Intell.*, 22, pp. 809-830, 2000.

[4] T. E. Boult, R. J. Micheals and X. Gao, "Into the Woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings," *Proc. IEEE*, vol. 89, no. 10, pp. 1382-1402, 2001.

[5] I. Pavlidis, V. Morellas and P. Tsiamyrtzis and S. Harp, "Urban surveillance systems: From the laboratory to the commercial world," *Proc. IEEE*, vol. 89, no. 10, pp. 1478-1497, 2001.

[6] J. Meniere, "System for monitoring a swimming pool to prevent drowning accidents," *US. Pat. No. 6,133,838*, October 2000.

[7] E. Menoud, "Alarm and monitoring device for the presumption of bodies in danger in a swimming pool," *US. Pat. No. 5,886,630*, March 1999.

[8] F. Guichard, J.M. Lavest, B. Liege, J. Meniere, "Poseidon Technologies: The world's first and only computer-aided drowning detection system," *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, DT-8 (Demo Session), 2001.

[9] A. Kam, W. Lu and W.-Y. Yau, "A video-based drowning detection system," *Springer-Verlag Lec. Notes. Comp. Sci. vol. 2353 (ECCV 2002)*, Part IV, pp. 297-311, 2002.

[10] S. Intille and J. Davis, "Real-time closed-world tracking" *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, pp. 697-703, 1997.

[11] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, pp. 679-698, 1986.

[12] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification*, John Wiley & Sons, 2001.

[13] C. S. Won, "A block-based MAP segmentation for image compression," *IEEE Trans. Circuits System Video Technology*, vol. 8, no. 5, pp. 592-601, 1998.

[14] F. Pia, "The RID factor as a cause of drowning," *Parks and Recreation*, 1994.

[15] K.-A. Toh and W.-Y. Yau, "Multi-modal biometrics fusion: Beyond optimal weighting," *Proc. the Seventh Intl. Conf. Automation, Robotics and Vision*, pp. 788-792, 2002. (Invited paper)

[16] K.-A. Toh, W.-Y. Yau and X. Jiang, "A reduced multivariate polynomials model for multi-modal biometrics and classifiers fusion," *IEEE Trans. Circuits Systems Video Technology (Special Issue on Image- and Video-Based Biometrics)*, 2003. (Accepted)

[17] T. -T. Lee and J. -T. Jeng, "The chebyshev-polynomial-based unified model neural networks for function approximation," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 28, no. 6, pp. 925-935, 1998.

[18] N. Ueda, "Optimal linear combination of neural networks for improving classification performance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 2, pp. 207–215, 2000.

[19] C. M. Bishop, *Neural networks for pattern recognition*, New York: Oxford University Press Inc., 1995.