# AUTOMATED CROWD DETECTION IN STADIUM ARENAS

Loris Nanni, [1] Sheryl Brahnam, [2] Stefano Ghidoni, [1] Emanuele Menegatti[1]
[1]DIE, University of Padua, Via Gradenigo, 6 - 35131- Padova – Italy e-mail: {loris.nanni, ghidoni, em}@dei.unipd.it
[2]CIS, Missouri State University, 901 S. National, Springfield, MO 65804, USA e-mail: sbrahnam@missouristate.edu

## ABSTRACT

In this paper, we present an approach for crowd detection based on an ensemble of classifiers which employ several feature representation schemes of crowd images, including, local ternary patterns, local binary patterns, and features based on the spatial gray level dependency matrix. A Support Vector Machine classifier is trained on each of these feature vectors. Classifier predictions are then combined by sum rule. Experiments are performed on a large dataset that contains challenging sequences of actual football matches recorded at a stadium arena. Experimental results confirm that the different feature representations give complementary information which is exploited by fusion rules. The method proposed in this paper is shown to outperform previous methods tested on the same dataset. MATLAB code using the different descriptors is available at http://www.dei.unipd.it/wdyn/?IDsezione=3314&IDgruppo_pass=124&preview=".

**Keywords**: crowd detection, local ternary patterns, spatial gray level dependency matrix, local binary patterns.

## 1. INTRODUCTION

Video surveillance is an active area of research. Technology has reached the stage where mounting cameras is cheap. Cities throughout the world continue to amass extensive networks of camera surveillance systems. This has led to something of a monitoring crisis. Despite the ample volume of available video data, little accountable information is being retrieved. Human monitoring is tiring, expensive, and ineffective. It is estimated, for instance, that monitoring 25 cameras 24/7 costs an average of $150K per year, and experiments at Sandia National Laboratories for the US Department of Energy showed that human attention to video monitors deteriorates to unacceptable levels after only 20 minutes of viewing [11]. A practical solution would automate the monitoring process, freeing personnel to further evaluate and respond to detected events. Some key technologies motivating research along these lines include video-based detection and tracking, video-based person identification, and large-scale surveillance systems [16]. The desirability of such research is reflected in governmental funding, such as the EU Chromatica and Prismatical programs, the U.S. Combat Zones [5] project, and the older U.S. program VSAM (Video Surveillance and Monitoring) [4] [18], which greatly promoted research in these areas as they relate to both the battlefield and the commercial sector [26] [13] [15].

A number of computer vision technologies that deal with automated video surveillance have also recently found significant commercial success [20] [28]. Much of the focus of this technology has been on building physical security applications, but other issues being addressed involve event detection and face recognition at a distance. Yet, despite the many recent advances, these state-of-the-art systems tend to fail at critical points when the environmental assumptions on which they were built no longer apply. One common environmental change that often results in system failure is crowding. Isolating events and actors in crowds is complicated by such things as background motion [19] and high levels of occlusion where only a small portion of the human shape is visible [30].

In recent years, algorithms for crowd detection have been gaining strong interest. There are several reasons motivating this research. Crowded environments are very difficult to monitor by human observers, whether live or via video surveillance, because the visual patterns are highly repetitive and the complexity of the movement characterizing the scene is often overwhelming. Moreover, crowds can form and grow quickly and unexpectedly turn violent. Crowd detection systems and the automated analysis of higher level crowd characteristics, such as crowd configuration [2], flow [29], and violence [14], hold out the promise of offering invaluable assistance to safety personnel in targeting areas of emerging threat. Crowd algorithms are also valuable because they enable intelligent video surveillance systems to analyze a wider range of scenarios, extending outside common intrusion detection and patrolling tasks [10], and crowd detection algorithms work well with existing camera networks, as the perimeter defining a crowd often spans distances lying outside the range of a single camera. Crowd detection is especially important in the context of intelligent and automated video surveillance systems intended for large venues and public events, such as football games and concerts, as well as for such common environments as city streets and underground train stations during peak hours.

The development of crowd detection algorithms is rather recent. One of the earliest works describing a system for detecting a crowd dates in the mid 90s [7]. In that early paper, the authors provided a high level description of crowds inspired by gas dynamics along with a low level machine vision solution. Other approaches developed over the years include methods based on motion analysis and texture analysis. In [23], for example, a motion based system is proposed that assumes particles are evenly spaced; these are then displaced according to optical flow. In [17], motion heat maps, together with a set of indicators for measuring motion entropy, are used to detect crowds. Finally, in [2], a motion-based system is proposed that detects the precise contours of a crowd from still images.

Texture analysis works especially well when the task is focused on measuring the entropy of an image rather than on finding objects characterized by precise shapes that can be geometrically described. In crowd detection, the main approach used in texture-based methods is based on evaluating the gray level dependency matrix (GLDM), a method of feature representation that dates back to the 1970s [12]. This feature set is very large. However, in algorithms using it for crowd detection, only a few features are retained. In [22], for example, four features (labeled contrast, homogeneity, energy, and entropy) are used as the inputs for a neural network that classifies crowd density. In [10], a novel set of features based on GLDM are proposed that provide a richer description of the co-occurrence matrix for analyzing smaller segments of images. Aside from GLDM, other systems have used the Histogram of Oriented Gradients (HOG) descriptor [9] and SIFT feature density [2]. Finally, of note is [21], where a number of texture-based methods are compared and evaluated for crowd detection.

In this paper, we present a novel texture-based approach for crowd detection using an ensemble of classifiers that employ several descriptors, or feature representation schemes, of crowd images, including, local ternary patterns (LTP), local binary patterns (LBP), and a feature set based on GLDM. A Support Vector Machine classifier is trained on each of these feature vectors. Classifier predictions are then combined by sum rule. Experiments are performed on a large dataset that contains challenging sequences recorded during real football matches at a stadium arena. In our experiments, a fusion approach obtains the best average result. Our experiments demonstrate that it is possible to develop crowd detection systems composed of different simple methods that perform competitively with more complex state-of-the-art systems. This is useful since combining *n* independent approaches lends itself easily to parallelization (for instance, a different descriptor can be given to each core in a system with a multicore processor) making the system suitable for real time applications without the need for hardcode optimization.

The remainder of this paper is organized as follows. In section 2, we provide a detailed description of the texture descriptors used in our experiments. In section 3, we outline and explain our proposed approach. In section 4, we describe the dataset and several experimental results for validating our approach. Finally, in section 4, we summarize our results and give suggestions for further research.

## 2. TEXTURE DESCRIPTORS

Automated crowd localization and classification is a difficult machine classification problem that we believe is best handled by combining multiple descriptors to boost performance. Good descriptors are invariant to image rotation and scale. In addition, they are robust in terms of variations in illumination. By combining descriptors, a system can utilize the best properties of each. The remainder of this section describes the various texture descriptors used in our proposed ensemble method.

### 2.1 Invariant Local Binary Patterns (LBP) [24]

LBP is an extensively studied local texture operator that possesses several excellent properties: low computational complexity, rotation invariance, and robustness in terms of illumination variations. LBP is a histogram that is based on a statistical operator that is calculated by examining the joint distribution of gray scale values of a circularly symmetric neighbor set of $P$ pixels around a pixel $\mathbf{x}$ on a circle of radius $R$. In this study we use a multiresolution descriptor that is obtained by concatenating two histograms calculated using the uniform bins with the following parameters: ($P$=8; $R$=1) and ($P$=16; $R$=2).

### 2.2 Local Ternary Patterns (LTP) [27]

A generalization of LBP is LTP, which represents gray-scale differences between the pixels using a ternary rather than a binary value. The difference between the gray value of a pixel $\mathbf{x}$ from the gray values in one of its neighborhood $\mathbf{u}$ is represented by three values, which are determined by the application of the threshold $\tau$ : 1 if $\mathbf{u} \geq \mathbf{x} + \tau$ ; -1 if $\mathbf{u} \leq \mathbf{x} - \tau$ ; else 0. This provides a more discriminant descriptor that is also less sensitive to noise. To reduce computational complexity, the ternary pattern is divided into two binary patterns by considering both the positive and the negative components. The histograms computed from these two patterns are then concatenated. In our system, two different parameter configurations are evaluated: ($P$=8; $R$=1) and ($P$=16; $R$=2).

### 2.3 Histogram of oriented gradients (HOG) [6]

HOG calculates intensity gradients from pixel to pixel and selects a corresponding histogram bin for each pixel based on the gradient direction. The HOG features extracted in our experiments use a 2×2 version of the HOG. The HOG features were extracted on a regular grid at steps of 8 pixels and stacked together considering sets of 2×2 neighbors to form a longer descriptor with more descriptive power.

### 2.4 Haralick texture features [12]

The Haralick texture features descriptor was proposed nearly 30 years ago to classify different categories of rock, but it is widely used today to classify many types of images. It is based on the SGLD, or the co-occurrence matrix.

Given an image with $N$ gray levels, the SGLD matrix at angle $\theta$ is a matrix of size $N \times N$. Each element in the matrix is a count of the total number of pairs of gray levels $i$ and $j$ at a distance $d$ along the direction $\theta$.

Thirteen features are calculated from a SGLD matrix at a fixed angle $\theta$: energy, correlation, inertia, entropy, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy, and two information measures of correlation (Implemented as in Haralick Texture Features Matlab Toolbox v0.1b www.bme.utexas.edu/reasearch/informatics). In this work we test the features set extracted using Haralick's method, which concatenates the features extracted by considering four angles (0°, 45°, 135° and 90°), with $d$=1.

### 2.5 Shape analysis [10]

The SGLD is capable of measuring texture by analyzing the transitions (i.e., the differences between the gray levels) between couples of pixels, and organizing them to form a histogram. Deeper studies reveal that this matrix contains a great deal of information that is only partially extracted by features commonly used in the literature [10]. For this reason, it is worth investigating novel features and methods in order to extract more information in a given framed scene.

The SGLD can be seen as a two-dimensional histogram that is created by setting up a grid of 256 x 256 locations (in the common case of 8-bit image depth), one for each grayscale value. Once the whole image has been scanned and each pixel couple considered, the SGLD represents how pixels change. If only smooth variations can be found in the image, the SGLD will be concentrated towards the diagonal, while abrupt changes will lead to peaks that have a certain distance from the diagonal. Information about where such transitions occur, however, is lost, since all contributions are summed up irrespective of the region they were observed.

To obtain a better characterization of SGLD, a set of new features was developed in [10], with the idea of describing the shape of the histogram in more detail. From a detailed shape description, it is then possible to obtain much more data than that provided by commonly used indicators.

One way to extract more information is to analyze the 3D shape of the histogram by considering several height values. Each level curve is then analyzed by approximating it with an ellipse. On each ellipse a number of parameters are measured, and, finally, the amount these parameters change over the different ellipses is measured. This leads to a number of useful indicators, the first one being the decrease of the axes of the ellipses, and how it fits with a linear model. The maximum and minimum of the axes is also valuable. Yet another parameter of interest measures the volume under the highest contour level, with its dual parameter (the volume of the SGLD over the same contour level) also considered. A further important parameter is the eccentricity of the ellipses, which measures the amount of strong pixel variations: an ellipse which is thin around the main diagonal indicates that the number of strong variations is negligible, and vice-versa. Further parameters that we consider in this work are the surface of the smallest ellipse that describes how smooth the upper part of the SGLD is, and its ratio with the widest ellipse. Finally, the number of SGLD locations that have zero height is measured.

## 3. PROPOSED APPROACH

In this paper we not only combine different texture descriptors but also different color constancy approaches (the color constancy approaches estimate the unknown light of a scene and try to normalize it to a standard light) as well as the contrast limited adaptive histogram equalization[1] (AH), using the function *adapthisteq.m* in MATLAB 7, as a preprocessing method to reduce the illumination problem [8]. The following color constancy approaches are tested: Grey-World (*GW*), max-RGB (*MR*), Shades of Grey (*SG*), and Grey-Edge (*GE*).

The system we propose in this paper classifies a given image in a given crowd density into two categories: *no crowd and low* (where less than 4 persons appear in an image) and *med and high crowd* (where 4 or more persons appear in the image). The number of persons in each image in our database was manually tabulated, and each image was labeled accordingly. Experiments were then performed that combined the different descriptors using the sum rule. In Table 1 we list the features tested in this paper.

| Descriptor | Short name | Dimension |
|---|---|---|
| Local Ternary Pattern | *LTP* | 604 |
| Local Binary pattern | *LBP* | 302 |
| Haralick texture features | *HAR* | 52 |
| Histogram of gradients | *HOG* | 81 |
| Shape analysis | *SA* | 36 |

**Table 1**. Descriptors used in the proposed system.

The original video resolution was set to 640 × 480 pixels (for more details on the dataset, see section 4.1). In our system, the images were divided into either four or sixteen regions. As seen in section 4, the performance of our best ensemble depends on how the full image is divided. Moreover, whereas the best stand-alone descriptor in both cases is LTP, the other approaches perform quite differently in the two cases.

## 4. EXPERIMENTAL RESULTS

The dataset chosen for training and testing the classifier was acquired in a small stadium during a football match. Inside the stadium, a number of Pan-Tilt-Zoom (PTZ) cameras, capable of framing every part of the venue, was installed. A subset of four cameras installed in different positions (one framing outside the venue) was then chosen as a source for the recordings, which started one hour before the match was scheduled to begin and ended half an hour after the match ended. The recorded images are the same images that the security officers observed while keeping the venue under control. In this way, we are assured the images include salient events since the actions of security professionals (they controlled the tilt/pan/zoom mechanisms of the cameras) determined which scenes were recorded.

Recorded sequences include scenes in which the public or the game appear alone, scenes where both are present in the same image, and other scenes that include people queuing up at a kiosk during the game break. Flows of people getting into and out of the stadium are also present, as well as scenes of the empty venue, which were useful for providing the classifier system some negative examples with high texture content.

[1] Function `adapthisteq.m` of the MATLAB 7

Unlike scenes framed by surveillance cameras in other environments, images taken in stadium environments have peculiarities that justify our choice of acquiring a new dataset instead of exploiting others already available, such as the one presented in [1]. Despite the large number of images that can be acquired in stadium environments, the framed scenes are highly repetitive. As a result, the number of different scene conditions is strongly reduced. This reduction is mainly due to the fact that security officers tend to leave cameras in the same position for extended periods of time and the fact that people infrequently move while intent on watching a game. Moreover, the public normally looks similar even when framed from different viewpoints. However, there are times when the natural light can shift dramatically during the course of recording, making the same scene appear very different.

Out of all the recordings, 901 frames were chosen to best represent all possible scenarios that were observed during the match (see Figure 1 and 2, for examples). The frames were then organized into 19 short subsequences. It should be noted that the sequences for training and testing the classifiers were revised in this paper with respect to what was done in [10]. In this work, special care was spent in assuring that the training and testing sets did not include sequences taken from the same camera with the same orientation (thus simulating a real application scenario). This choice lead to a harder test for the classifier generalization capabilities, and justifies the lower performance obtained by the system described here with respect to [10].

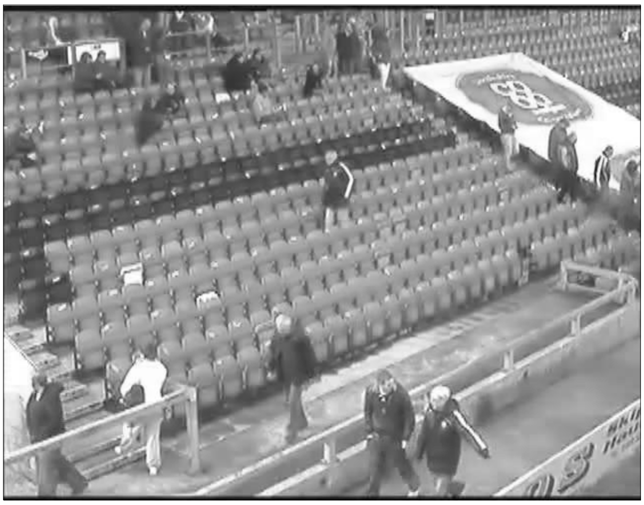

Figure 1. A typical image acquired inside a large venue.
The crowd is not the only framed entity providing high texture content.

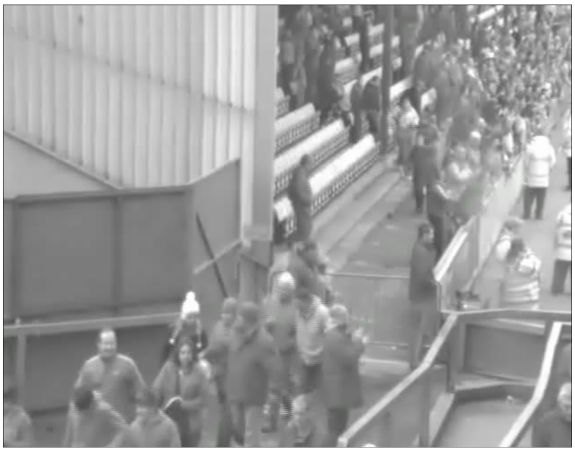

Figure 2. Crowd can be observed from different scales in the same frame.
Other objects with strong texture content are also present in the image.

### 4.2 Experiments

The area under the Receiver Operating Characteristic curve (AUC) is used as the performance indicator. The area under the ROC is considered one of the most reliable performance indicators [25] as it is based on both sensitivity and specificity. The ROC curve is a plot of the sensitivity versus false positives (1 - specificity). The error area under the ROC curve (EUC) can be interpreted as the probability that the classifier will assign a lower score to a randomly chosen positive sample than to a randomly chosen negative sample.

As the testing protocol, we used the leave-one-out cross-validation method, where in each fold the frames of a given sequence are used as a testing set, while the other frames of the other eighteen sequences are used to build the training set.

Table 2 reports the results in the first experiment, where we compare the different pre-processing approaches (adaptive histogram equalization and the four color constancy approaches) using the most widely used descriptor for person identification, namely, the HOG descriptor. The label NO is the performance obtained when no preprocessing approach is applied.

| **Pre-Processing** | ***4 regions*** | ***16 regions*** |
|---|---|---|
| *NO* | 67.38 | 77.30 |
| *AH* | **76.02** | **81.27** |
| *GW* | 68.02 | 77.41 |
| *MR* | 68.13 | 77.15 |
| *SG* | 68.75 | 77.50 |
| *GE* | 76.24 | 77.94 |

Table 2. Accuracy obtained using different feature sets.

As can be seen in Table 2, it is clear that color constancy has little impact, while the application of adaptive histogram equalization (AH) improves performance.

| **Pre-Processing** | ***4 regions*** | ***16 regions*** |
|---|---|---|
| *HOG* | 76.0 | 81.3 |
| *HARA* | 60.4 | 82.5 |
| *LTP* | 82.4 | 93.0 |
| *LBP* | 75.1 | 91.9 |
| *SA* | 73.1 | 72.2 |
| *FUS_4* | **83.2** | 92.5 |
| *FUS_16* | 81.8 | **93.2** |
| *FUS* | 82.5 | 93.1 |

**Table 3**. Comparison among different approaches.

In table 3 we compare the descriptors (preprocessed by AH) described in section 2. We also report the following fusion approaches:

- *FUS_4* is the weighted fusion between HOG (weight 1) and LTP (weight 4); it obtains the best performance in the *4 regions* testing protocol;
- *FUS_16* is the weighted fusion between LBP (weight 1) and LTP (weight 3); it obtains the best performance in the *16 regions* testing protocol;

- *FUS* is the weighted fusion of HOG (weight 1), LBP (weight 2), and LTP (weight 6).

The following conclusions can be drawn from the results reported in Table 3:

- LTP outperforms the other approaches;
- LBP obtains performance similar to LTP in the *16 regions* set (outperforming the other tested descriptors), while in the *4 regions* set it obtains a performance similar to SA and HOG;
- The fusions are quite useful but perform differently than they do in other problem domains (e.g., image sub-cellular classification [3]); it does not enhance the performance of the best single descriptor.

In Figure 3 we report an example where only one subwindow (surrounded by a black frame) is misclassified by our proposed method. Notice that the misclassified example contains three persons, so it is quite similar to the class "*med and high crowd*" that is composed by the images containing four or more persons.

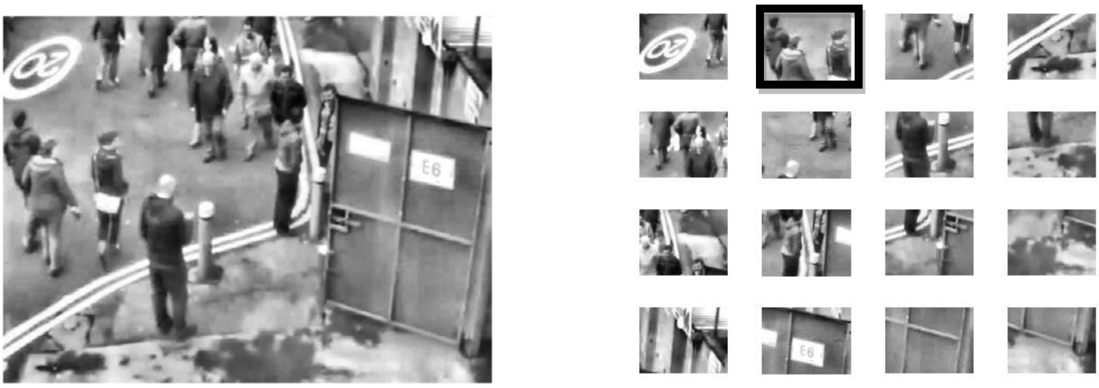

Figure 3. Example classified best by our system.

## 5. CONCLUSION

This paper focused on the study of texture descriptors for training an ensemble of machine learning algorithms for crowd image classification. The system proposed in this work is tested on a difficult dataset built using video sequences of real football matches at a local stadium arena. We performed several tests using sequences extracted from different cameras using two different testing protocols based on the original images, size 640×480, being divided into either four or sixteen regions. In the first protocol, the systems are trained and tested with the full image divided into four regions; in the second protocol the systems are trained and tested with the full image divided into sixteen regions.

Based on an analysis of prior research in other domains, we propose a method for automating crowd localization based on a set of SVMs trained using different descriptors. For combining the different descriptors, we train a SVM separately for each descriptor with the results combined using a weighted sum rule. It is interesting to note that the behavior of the each approach changes depending on the two different testing protocols.

In future work we plan on improving the performance of our system by evaluating other texture descriptors and different methods for combining ensemble evaluations. Moreover, since the images are acquired by cameras that are rarely moved, it may also be possible to develop an approach based on background subtraction. Our idea is that the model for the background could be rebuilt after every camera motion.

## REFERENCES

[1] Ali, S., and Shah, M., “A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis,” in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[2] Arandjelovic, O., “Crowd detection from still images,” in British Machine Vision Conference (BMVC), 2008.

[3] Chebira, A., Barbotin, Y., Jackson, C., Merryman, T., Srinivasa, G., Murphy, R. F., and J., J. K., “A multiresolution approach to automated classification of protein subcellular location images,” *BMC Bioinformatics,* vol. 8, pp. 210, 2007.

[4] Collins, R., Lipton, A. J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., and Wixson, L., *A system for video surveillance and monitoring*, The Robotics Institute, Carnegie Mellon University, Pittsburgh PA, 2000.

[5] Combat Zones That See U.S. Government DARPA Project.

[6] Dalal, N., and Triggs, B., “Histograms of oriented gradients for human detection,” in 9th European Conference on Computer Vision, San Diego, CA, 2005.

[7] Davies, A. C., Yin, J. H., and Velastin, S. A., “Crowd monitoring using image processing,” *Electronics Communication Engineering Journal,* vol. 7, no. 1, pp. 37-47, 1995.

[8] Eustice, R., Pizarro, O., Singh, H., and Howland, J., “UWIT: Underwater image toolbox for optical image processing and mosaicking in MATLAB,” in International Symposium on Underwater Technology, Tokyo, Japan, 2002, pp. 141-145.

[9] Gárate, C., Bilinsky, P., and Bremond, F., “Crowd event recognition using hog tracker,” in Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009, pp. 1-6.

[10] Ghidoni, S., Cielniak, G., and Menegatti, E., “Texture-based crowd detection and localisation,” in International Conference on Intelligent Autonomous Systems (IAS-12), 2012.

[11] Haering, N., Venetianer, P. L., and Lipton, A., “The evolution of video surveillance: An overview,” *Machine Vision and Applications,* vol. 19, pp. 279–290, 2008.

[12] Haralick, R. M., “Statistical and structural approaches to texture,” *Proceedings of the IEEE,* vol. 67, no. 5, pp. 786-804, 1979.

[13] Haritaoglu, I., Harwood, D., and Davis, L., “Real time surveillance of people and their activities,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 809–830, 2000.

[14] Hassner, T., Itcher, Y., and Kliper-Gross, O., “Violent flows: Real-time detection of violent crowd behavior,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 1-6.

[15] Horprasert, T., Harwood, D., and Davis, L., “A statistical approach for real-time robust background subtraction and shadow detectio,” in IEEE Frame-Rate Workshop, Kerkyra, Greece, 1999.

[16] Hu, W., Tan, T., Wang, L., and Maybank, S., “ A survey on visual surveillance of object motion and behaviors,” *IEEE Transactions on Systems, Man, and Cybernetics Part C,* vol. 34, no. 3, pp. 334-352, 2004.

[17] Ihaddadene, N., and Djeraba, C., “Real-time crowd motion analysis,” in 19th International Conference on Pattern Recognition (ICPR 2008), 2008, pp. 1-4.

[18] Kanade, T., Collins, R., Lipton, A., Anandan, P., and Burt., P., “Cooperative multisensor video surveillance,” in 1997 DARPA Image Understanding Workshop, 1997, pp. 3-10.

[19] Ke, Y., Sukthankar, R., and Hebert, M., "Volumetric features for video event detection," *International Journal of Computer Vision,* vol. 88, no. 3, pp. 339-362, 2010.

[20] Lipton, A., Heartwell, C., Haering, N., and Madden, D., "Automated video protection, monitoring & detection," *IEEE Aerospace and Electronic Systems Magazine,* vol. 18, no. 5, pp. 3-18, 2003.

[21] Marana, A. N., Costa, L. F., Lotufo, R. A., and Velastin, S. A., "On the efficacy of texture analysis for crowd monitoring," in International Symposium on Computer Graphics, Image Processing, and Vision (SIBGRAPI '98), 1998, pp. 354-361.

[22] Marana, A. N., Velastin, S. A., Costa, L. F., and Lotufo, R. A., "Estimation of crowd density using image processing," *Image Processing for Security Applications,* no. Digest no.: 1997/074, pp. 11/1–11/8, 1997.

[23] Mehran, R., Oyama, A., and Shah, M., "Abnormal crowd behavior detection using social force model," in IEEE Computer Vision and Pattern Recognition (CVPR 2009), 2009, pp. 935-942.

[24] Ojala, T., Pietikainen, M., and Maeenpaa, T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Ieee transactions on pattern analysis and machine intelligence,* vol. 24, no. 7, pp. 971-987, 2002.

[25] Qin, Z. C., "ROC analysis for predictions made by probabilistic classifiers," in Fourth International Conference on Machine Learning and Cybernetics, 2006, pp. 3119-312.

[26] Stauffer, G., "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 747-757, 2000.

[27] Tan, X., and Triggs, B., "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *Analysis and Modelling of Faces and Gestures,* vol. LNCS 4778, pp. 168-182, 2007.

[28] Tian, Y.-l., Brown, L., ·, A. H., Lu, M., Senior, A., and Shu, C.-f., "IBM smart surveillance system (S3): event based video surveillance system with an open and extensible framework," *Machine Vision and Applications,* vol. 19, pp. 315-327, 2008.

[29] Wang, B., Ye, M., Li, X., Zhao, F., and Ding, J., "Abnormal crowd behavior detection using high-frequency and spatio-temporal features," *Machine Vision & Applications,* vol. 23, no. 3, pp. 501-511, 2012.

[30] Zhan, B., Monekosso, D. N., Remagnino, P., Velastin, S. A., and Xu, L.-Q., "Crowd analysis: A survey," *Machine Vision and Applications,* vol. 19, pp. 5-6, 2008.