An Empirical Study for Human Behavior Analysis

Jia Lu, Auckland University of Technology, Auckland, New Zealand

Jun Shen, Auckland University of Technology, Auckland, New Zealand

Wei Qi Yan, Auckland University of Technology, Auckland, New Zealand

Boris Bačić, Auckland University of Technology, Auckland, New Zealand

ABSTRACT

This paper presents an empirical study for human behavior analysis based on three distinct feature extraction techniques: Histograms of Oriented Gradients (HOG), Local Binary Pattern (LBP) and Scale Invariant Local Ternary Pattern (SILTP). The utilised public videos representing spatio-temporal problem area of investigation include INRIA person detection and Weizmann pedestrian activity datasets. For INRIA dataset, both LBP and HOG were able to eliminate redundant video data and show human-intelligible feature visualisation of extracted features required for classification tasks. However, for Weizmann dataset only HOG feature extraction was found to work well with classifying five selected activities/exercises (walking, running, skipping, jumping and jacking).

KEYWORDS

Histograms of Oriented Gradients (HOG), Human Behavior Recognition, Local Binary Pattern (LBP)

INTRODUCTION

Rapid increase of surveillance cameras capturing human activities, processing of events of interest in a scene is still considered as an on-going challenge. In general, for loosely defend 'intelligent' surveillance, a video acquired from networked cameras should be processed in such way that it could provide meaningful information that we could use and further process if needed.

As an active research area in intelligent surveillance, event recognition mining and reasoning can improve the accessibility and reusability for a large number of media collections (Maryam & Reza, 2012). Event recognition, which has a wide range of applications, can also be utilized to identify particular events as a function to find out abnormal human behaviours (Chen & Zhang, 2006). Event in our real world can be defined as occurrence happened in a determinable space and time (Popoola & Wang, 2012).

To reduce human labour involved in visual detection and events/activities recognition from video streams and to label a surveillance system as 'intelligent' at present time, it is necessary to achieve a degree of automation in information processing from videos. To achieve these objectives, we aim to design computer infrastructures for monitoring our environments in 24 hours per day, seven days one week.

DOI: 10.4018/IJDCF.2017070102

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Despite growing number of surveillance videos from various sources (analogue and digital) that have been analysed for decades, object, event, activity detection and recognition have been well investigated, this problem still remains. In recent years, there is also growing attention in our research community for data processing from the surveillance with focus on how to detect and recognize objects with spatio-temporal relationships (Krumm, et al., 2000) (Siebel & Maybank, 2002). The importance for this research is at semantic understanding of human behaviour. Automated human behaviour recognition plays a pivotal role in storing video streams into a database for further analysis. In literature, there are two categories of automated event analysis, namely spatio-temporal models and parametric models (Rui & Anandan, 2000) (Cutler & Davis, 2000).

In this paper, we assert that in surveillance systems there is no single best approach. In addition, the process of Feature Extraction (FE) as video data reduction and intermediate visualization of data processing stages may provide additional insight into specific surveillance problem or broader context that may inform further advancements in this field.

This paper is organized as follows: Section 2 includes background and related work of person and event detection in surveillance, ANN and feature selection techniques; Section 3 will describe utilized datasets, application of existing feature extraction techniques with comments on perceived visualisation of intermediate data processing for classification tasks; Section 4 reports on the results achieved; and the conclusions and future work will be summarised in Section 5.

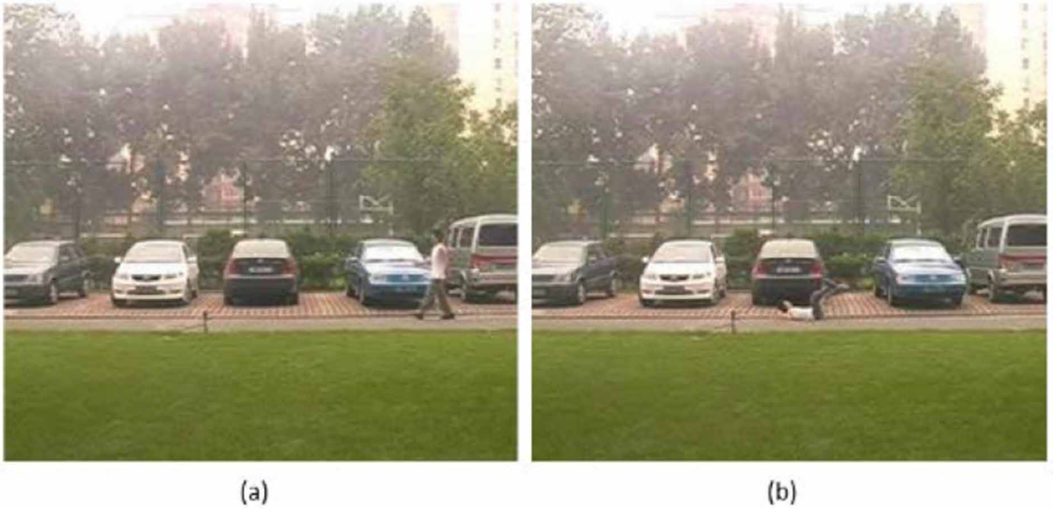
BACKGROUND AND RELATED WORK

In surveillance, we describe an event in six facets, namely, What, When, Who, Why, Where and How (5W1H) that could be generalized to feature any surveillance events (Westermann & Jain, 2007). Aligned with studies in video mining and video retrieval (Dai, Zhang, & Li, 2006) (Geetha & Narayanan, 2010), events are regarded to consist of these six 5W1H major components in event recognition and modelling (Xie, Sundaram, & Campbell, 2008). In computing, visual event represents an action or occurrence that could be quantified and recognised by (computing) machine. Similarly, the definition of an event in this paper is the occurrence of something at a particular time and at specific location. In order to facilitate a computer to record, index and arrange video events for users' post-analysis, the events have a number of attributes including ID, time, location and description. According to the attributes, an event is detected and classified into different classes from the videos in surveillance. Based on categories of an event, we can group the detected events as normal and abnormal ones. For example, Figure 1 shows a normal and abnormal event. Normally, a pedestrian should walk in standing position as in Figure 1(a) or when the walker/bystander falls down as in Figure 1(b) the abnormal event should be detected, and a surveillance alarm should be generated correspondingly and automatically.

Spatial-temporal model and periodic model (Rui & Anandan, 2000) (Cutler & Davis, 2000) are proposed to detect and analyse periodic motions. Video mining is still playing a key role in development of next-generation of video search capabilities (Xie & Yan, 2009). At present, the problem that still remains for video event search is lack of effective indicators to describe the content of video data. In addition to four general phases of event mining (Valera & Velastin, 2005), it is also important to extract existing semantic patterns (Maryam & Reza, 2012).

Moreover, event recognition can be split into twofold, which encompass model-based approaches and appearance-based approaches. In the first approaches, Bayesian networks typically have been used to recognize the simple events or static postures from video frames (Intille & Bobick, 2001), meanwhile Hidden Markov Model (HMM) also has been applied to human behaviour recognition (Oliver, Rosario, & Pentland, 2000) (Tran & Davis, 2008). Appearance-based approaches are based on salient regions of local variations in both spatial and temporal dimensions (Laptev, 2005) (Niebles, Wang, & Fei-Fei, 2008). Boosting is adopted to learn for a cascade of filters for efficient visual event detection (Ke, Sukthankar, & Hebert, 2005). In addition, grammar-based and statistical-based methods

Figure 1. Examples of surveillance events: (a) A normal event; (b) An abnormal event (e.g. a person may have fallen in front of a car in the centre)



(Naphade & Huang, 2002) could also be categorized by dimension of sampling support, characteristics and mathematical modelling. Moreover, the supportive samples can be a pixel, a region or a frame of the abnormality (Tziakos, Cavallaro, & Xu, 2010). Relevance Feedback (RF) was introduced to retrieve results of a specific query by enquiring subjective user's opinion incorporated in the learning process (Su, Zhang, Li, & Ma, 2003). Based on multi-camera surveillance data, it was found that event description language is possible for annotating events (Velipasalar, Brown, & Hampapur, 2006). An unsupervised model was proposed in event recognition (Xie, Sundaram, & Campbell, 2008), which consists with outlier identification and model adaptation.

ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANN) are maintaining their popularity especially with the recent advancements in deep learning (Bajpai, Jain, & Jain, 2011). In machine learning and data mining, ANN approaches cover broad areas such as data analysis, clustering and pattern recognition. A neuron model (connectionist model) is proposed in which computations could be processed by a network of simple binary neurons (Sondak & Sondak, 1989). Most ANN algorithms are based on supervised learning model. One of the examples is Back Propagation (BP) algorithm, which is common for multilayer feed-forward network, consisting of three (or more) layers and required training data to build the model. The traditional BP algorithm makes the input/output problem converted into a non-linear optimisation problem by using iterative negative gradient descent algorithm (delta rule). BP algorithm is one of the commonly used multilayer networks, relying on supervised learning. As a result of supervised learning, neural network converges into a local minimum and also exhibits slow learning speed associated with computational complexity (Burse, Manoria, & Kirar, 2011).

ANN can be seen as a parallel system, which is similar to human brain. Similar to our brain, neurons are the basic processing units in various ANN architectures. However, the ANNs and connectionist systems in general are also unlike a real human brain given their implementation variations for signal processing (e.g. neuron activation function), weighted interconnectivity (as numerical equivalent to synapses) and input/output processing data values or amplitudes (Bajpai, Jain, & Jain, 2011). In addition, a classifier/connectionist system may be implemented as a clustering method or as a decision tree. More importantly, for traditional ANN, after the data-training phase is completed, their internal structure cannot be changed and any new data added into re-training the model may result in phenomenon known as 'catastrophic forgetting' (Miller & Khan, 2011). Moreover, components of a human detection system are with equivalent ANN models (ANN, SVM and AdaBoost) and accompanying experimental studies (Enzweiler & Gavrila, 2008).

FEATURE EXTRACTION TECHNIQUES

In this paper, the objective of feature extraction techniques (FET) in traditional ANN approaches is to pre-process and transform data in such way to eliminate redundant information and produce feature space that discriminates well patterns that are robust to changes in particular context. For the scope of this study, it is also expected that as a result of pre-processing involved in FET, it would be possible to produce human intelligible visual representation of non-redundant data representing patterns. One of the FET that is known to work well with static images, textures and also to be robust to rotation variations is named as Local Binary Pattern (LBP). Over years, LPB method has become well established in video and image processing due to its relative simplicity and discriminative power of pattern representation (Wang & He, 1990). In addition to LPB, Histograms of Oriented Gradients (HOG) is another popular FET, which is adopted for object recognition in computer vision suitable for human behaviour detection and recognition (Mohan, Papageorgiou, & Poggio, 2001).

The HOG feature extraction is one of the important steps for people detection and recognition (Dalal & Triggs, 2005). The HOG features for human detection have been trained and tested after normalization, gradient computation and spatial organization. The main idea of HOG descriptor is to calculate the occurrences of gradient orientation in localized portions of an image.

Aligned with prior work in event analysis (Zelnik-Manor & Irani, 2001), to analyse video events in this study, the first task was to obtain experimental video data and predefine known tested and ignored behaviours. Aligned with literature review to achieve the objective of analysing human activity, the most important step aims to extract the visual features from video frames. Based on our human body and region of interest, the follow-up investigation is to extract possible features, which can be employed to automatically describe the human behaviour via classification tasks. For classification, we have chosen *k*-nearest neighbours (KNN), multi-layer perceptron (MLP) and decision tree (C4.5 algorithm).

EXPERIMENTAL SETTINGS

At present, there are publicly available video datasets utilized for human behaviour recognition. In this study, we used the publicly available datasets INRIA for pedestrian detection and Weizmann datasets for human behaviour recognition.

In the INRIA dataset there are 614 images and totally 2416 pedestrians. The pedestrians are all standing, viewing angles are various including front, back and various sides. The other 1218 images are non-pedestrian. The negative examples will be input into the neural network for training the neural network. Table 1. shows ten output classes associated with event labels of the Weizmann dataset which contains 90 video clips with the resolution of 180×144. The videos show nine participants in total.

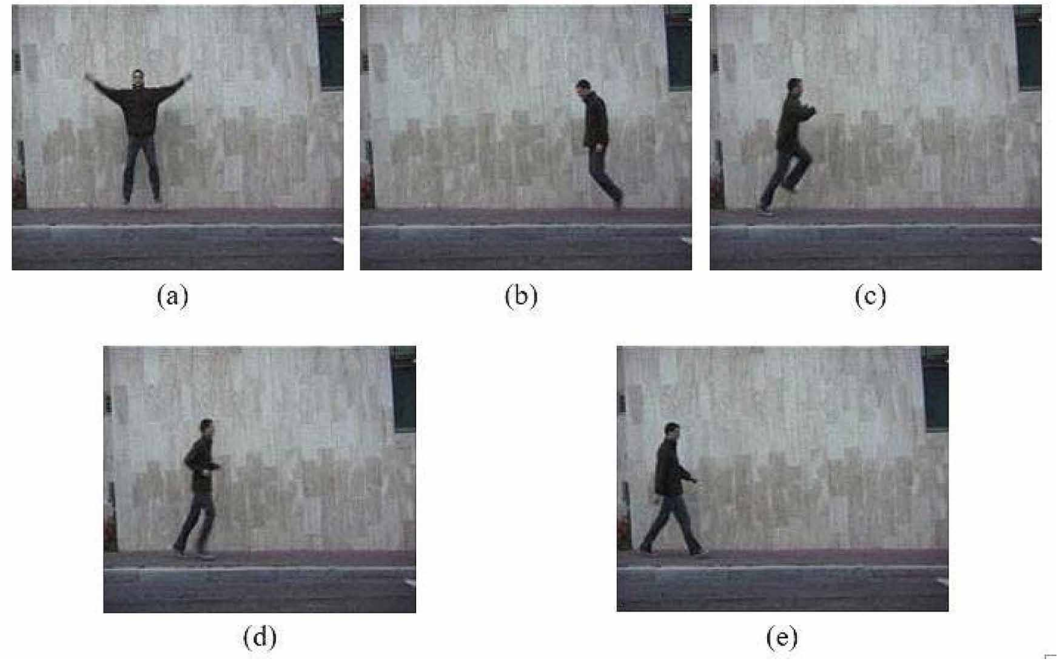
Each class contains at least nine video clips captured by a static camera view, with static representation of viewed outdoor environment. Figure 2 shows an example of video frames in Weizmann datasets which are adopted for this research project. There are totally five images in this example; each image represents one behavior of a person.

To design and implement human behavior recognition is the main purpose of this paper. As shown in Figure 3, the flowchart of human behavior recognition for each step is clearly pointed out through six different stages as the structure of this research.

Table 1. The Weizmann Dataset: Output classes and event labels

Class	Event Label	Class	Event Labels
1	Walk	6	One-hand wave
2	Run	7	Two-hands wave
3	Jump	8	Jump in place
4	Gallop sideways	9	Jumping jack
5	Bend	10	Skip

Figure 2. Video frame example: (a) Jacking; (b) Jumping; (c) Skipping; (d) Running; and (e) Walking



The first four steps are thought as the low-level processing which mainly includes detecting the Region of Interest (ROI), segmenting ROI region and extracting computable features. Three distinct Feature Extraction Techniques (FET) were investigated and implemented in human behavior recognition in order to achieve the objective of this paper, which contains Histogram of Oriented Gradient (HOG), Local Binary Pattern (LBP) and Scale Invariant Local Ternary Pattern (SILTP).

To improve the performance, local histogram is normalized by calculating the intensity in a large region across the image called ‘block’, then is used to normalize all cells within the block. After the normalization, illumination and shadows have been greatly changed. Figure 4 shows the grey scale images from Weizmann dataset. Histograms of Oriented Gradients (HOG) descriptors have the advantages, which effectively describe local shape of the images. Figure 5 shows the extracted HOG features of different human behaviours, and the length of HOG feature vector in the experiment is 1440. Figure 6 illustrates an example of the deployment of HOG feature vectors, one block (B_1) contains four cells (C_{11}, C_{12}, C_{21}, C_{22}). The second block B_2 also contains four cells (C_{21}, C_{22}, C_{31}, C_{32}). Therefore, HOG descriptor in this experiment are organized as eight-by-eight cell size. The eight-

Figure 3. The flowchart of the steps of human behavior recognition

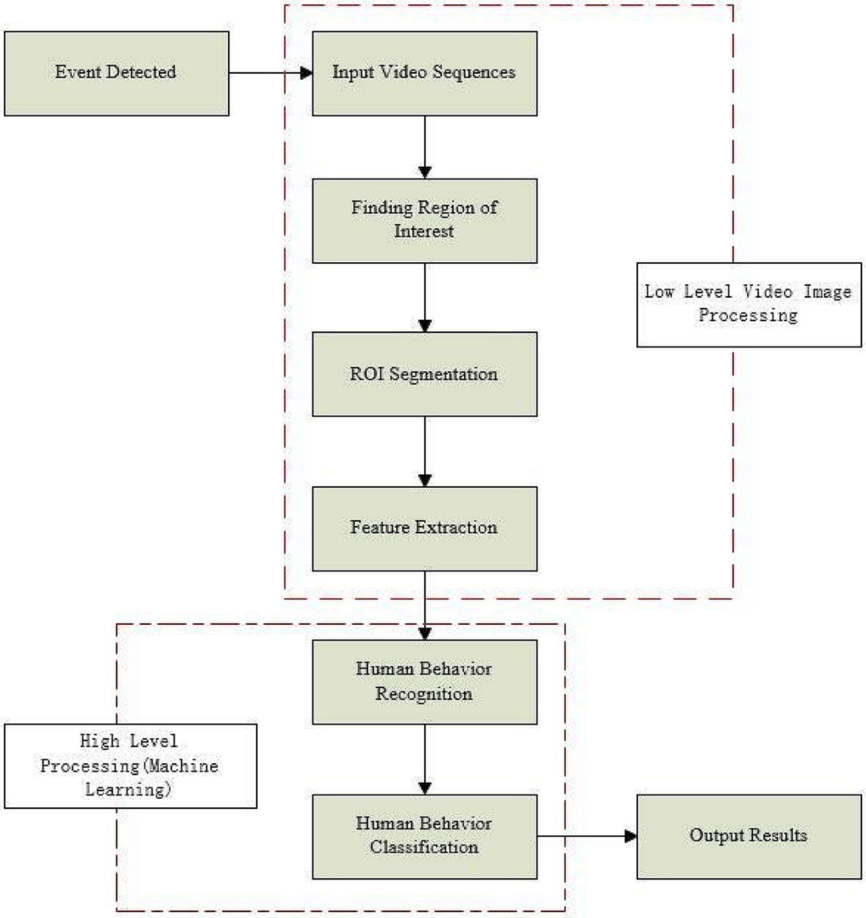

Event Detected
Input Video Sequences
Finding Region of Interest
ROI Segmentation
Feature Extraction
Low Level Video Image Processing
Human Behavior Recognition
High Level Processing(Machine Learning)
Human Behavior Classification
Output Results

Figure 4. Video frame processing example: (a) Background and foreground separation is followed; (b) Grey scale image conversion

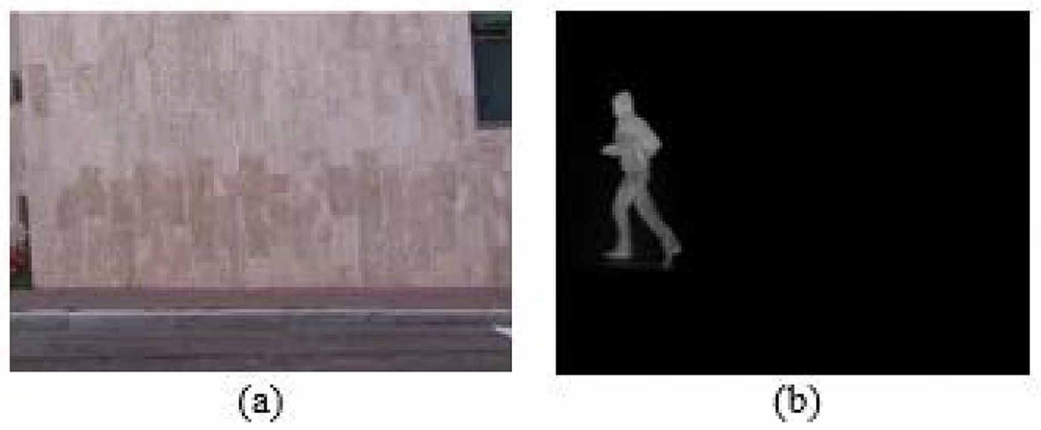


Figure 5. Feature extraction by using HOG for human behaviours: (a) Jacking; (b) Jumping; (c) Running; (d) Skipping; and (e) Walking

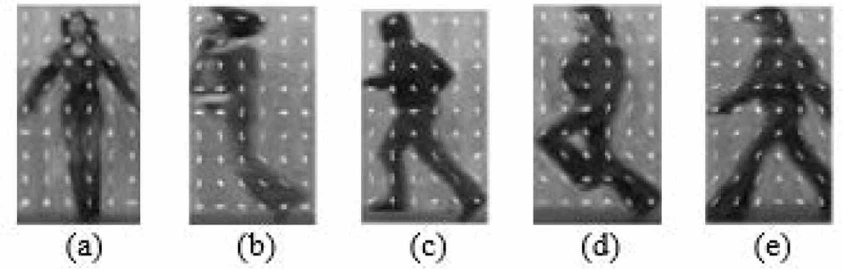
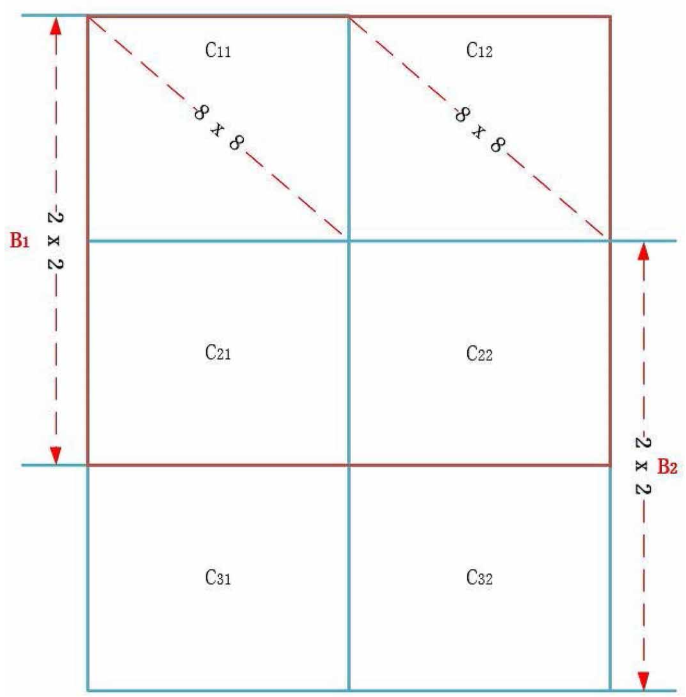


Figure 6. The deployment of HOG feature vectors

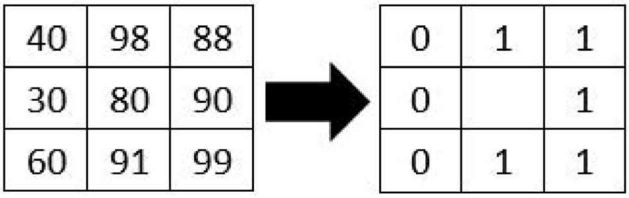


by-eight cell size does not include too much shape information. However, cell size will increase the dimensionality of the HOG feature vectors which also increase the computational time.

Both visualisation outcomes of HOG processing (Figure 3.3 and Figure 3.4) are human intelligible and have potential to inform and inspire modelling decisions in surveillance. HOG descriptors have the advantages that effectively describe local shape of the images. By changing number of bins and cell size of the histogram, it is able to capture an image of the local region.

Local Binary Pattern (LBP) is an operator to describe the local texture feature of an image. The advantage of this operator is the characteristics of rotational invariance and grey scale invariance. Local Binary Pattern (LBP) is a simple and effective algorithm for feature extraction as LBP describes a relationship between a pixel and the others around it as shown in Figure 7. The extraction of LBP depends on the intensity of a grey scale pixel. For one target pixel, there are at more than 8 pixels around it, if the intensity of surrounding pixels is greater than that of the pixel in the middle, they

Figure 7. The relationship between pixels in LBP



40	98	88
30	80	90
60	91	99

0	1	1
0		1
0	1	1

will be marked by 1 or else by 0 respectively. The eight-bit binary number will be the LBP feature of the pixel in the central. For the LBP feature extraction in practice, the original image will be divided into a number of cells with a fixed size (16 × 16). After that, we applied this method to obtain the eight-bit binary number and then calculate the histogram for each cell and applied normalization to the histograms. Finally, we combine all these histograms to get a feature vector of the whole image. Figure 8 shows experimental results of LBP feature extraction. The feature vectors can be input into a neural network for training and testing.

Compared to HOG processing visualisation (Figure 4 and Figure 5), LBP processing visualisation is perceived as inferior and not always human-intelligible.

Like Local Binary Pattern (LBP), Scale Invariant Local Ternary Pattern (SILTP) as another local pattern was studied in 2010 which is effective for illumination variations. SILTP is very similar to LBP, grayscale intensity of the central pixel will be compared with its neighbour pixels. However, a scale factor was proposed in SILTP to indicate the range and one more comparison will be conducted in SILTP. Figure 9 illustrates the simple mechanism of the SILTP operator. Once the grayscale intensity

Figure 8. Feature extraction by using LBP for different images. In row (b) only the first two images allow observer to interpret person's presence.

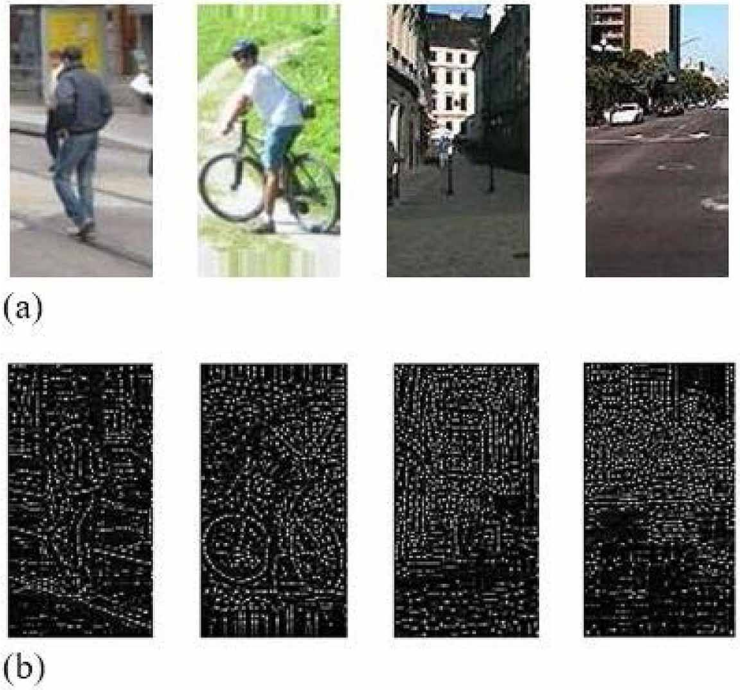
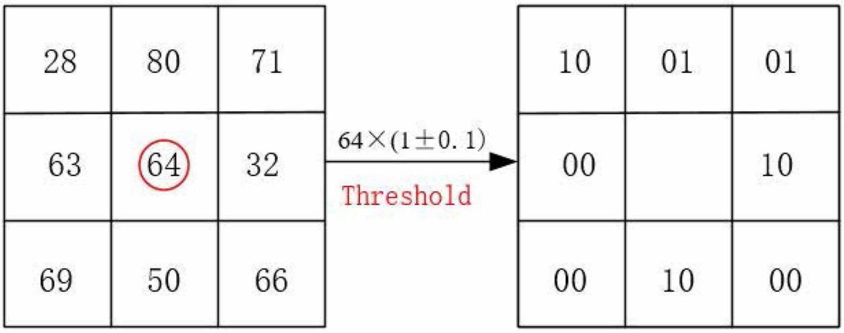


Figure 9. The simple mechanism of SILTP operator



28	80	71
63	64	32
69	50	66

10	01	01
00		10
00	10	00

of one pixel is less than the range of the minimum value of the central pixel, the pixel will be marked as “10”. If the intensity value of one pixel is great than the range of the maxima value of the central pixel, the pixel will be marked as “01”. If the grayscale value of one pixel is within the range of central pixel 64 × (1±0.1), the pixel will be marked as “00”.

Scale Invariant Local Ternary Pattern (SILTP) has three advantages in the local image including efficiency in computation, robustness to local noises and robustness to illumination changes. The length of a SILTP feature vector in the experiment is 3648. Figure 10 shows the extracted output of SILTP features for different human behavior analysis based on the Weizmann dataset.

RESULTS AND ANALYSIS

There are several of standard evaluations in information retrieval which are calculated from the obtained confusion matrix including accuracy, recall and precision.

Figure 11 and Figure 12 show the results of pedestrian detection using the INRIA dataset. In the top-left corner of the images, the red marker labelled ‘P’ represents that the object is pedestrian and the red marker labelled ‘B’ refers to non-pedestrian. According to the resultant comparisons between HOG and LBP with different classifiers in Table 2, the overall performance of HOG is more accurate than that of LBP, which proved HOG is more suitable in this experiment, even LBP is faster than HOG due to the less intensive computation involved in feature extraction.

Figure 10. Feature extractions by using SILTP for different human behaviors: (a) Jacking; (b) Jumping; (c) Running; (d) Skipping; and (e) Walking

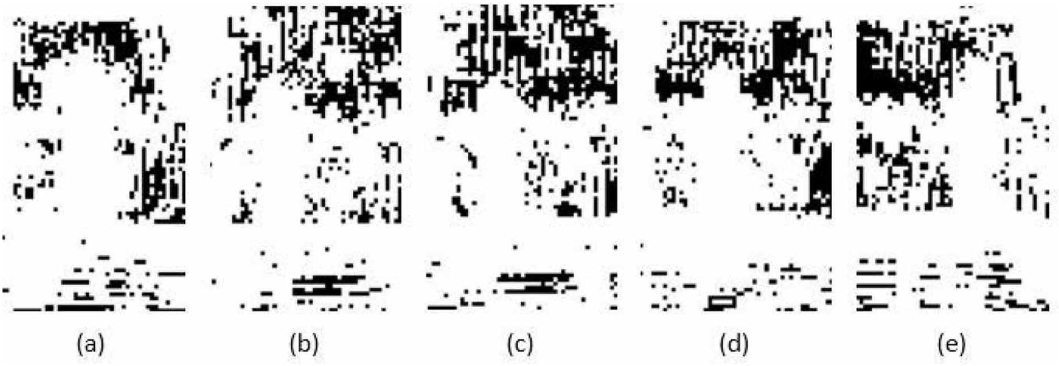


Figure 11. The testing images for pedestrian detection using HOG

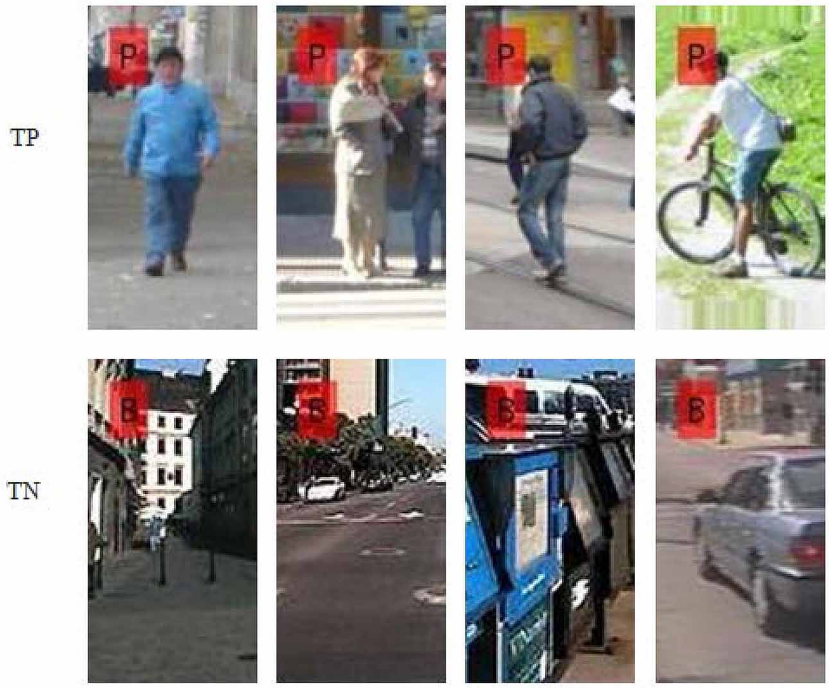


Figure 12. The testing images for pedestrian detection using LBP

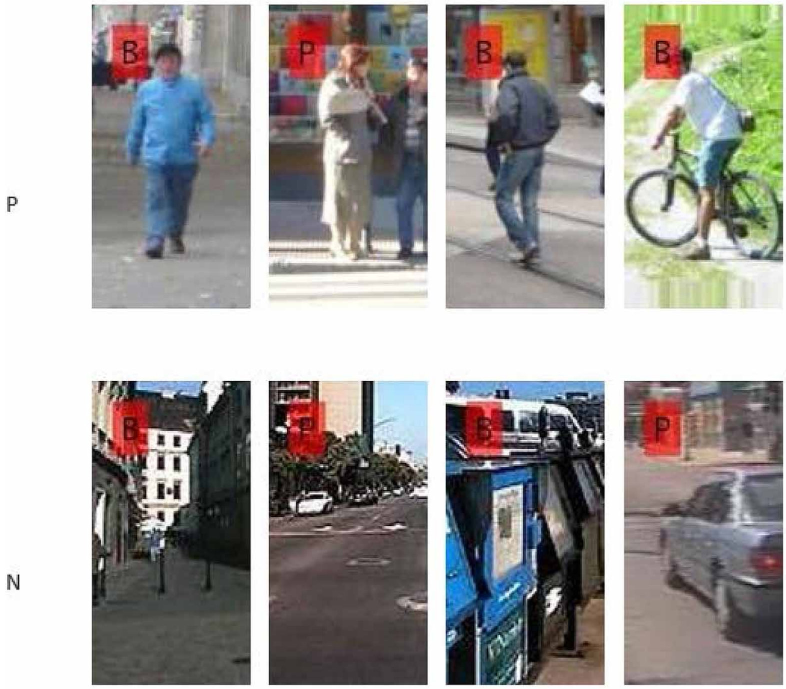


Table 2. Accuracy comparisons between features HOG and LBP using diverse classifiers

Classifier	HOG		LBP	
	Pedestrians	Non-Pedestrians	Pedestrians	Non-Pedestrians
KNN	94.2%	94.7%	94.9%	94.1%
Decision Tree	87.9%	87.7%	86.4%	86.8%
MLP	95.6%	94.4%	93.9%	93.0%

Figure 13 and Figure 14 show the results of pedestrian detection in various scenes. When a moving object is detected, the feature of this object will be extracted and sent to a neural network for training and testing so as to implement the recognition. The objects marked with green rectangle are the classified pedestrian and the red rectangle marked objects reflects the moving objects but not relevant to pedestrians.

Figure 13. The results for pedestrian detection (marked with red 'P' marker) and non-pedestrian (marked with red 'B' marker) with external testing images using HOG

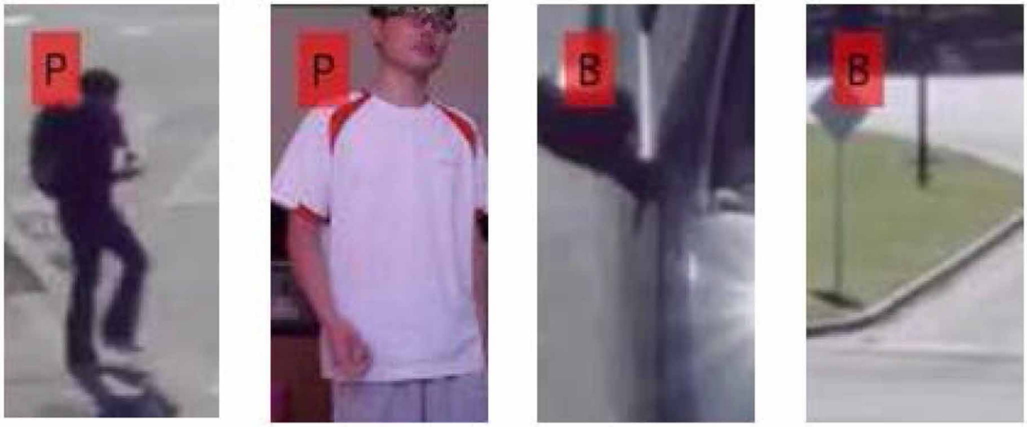
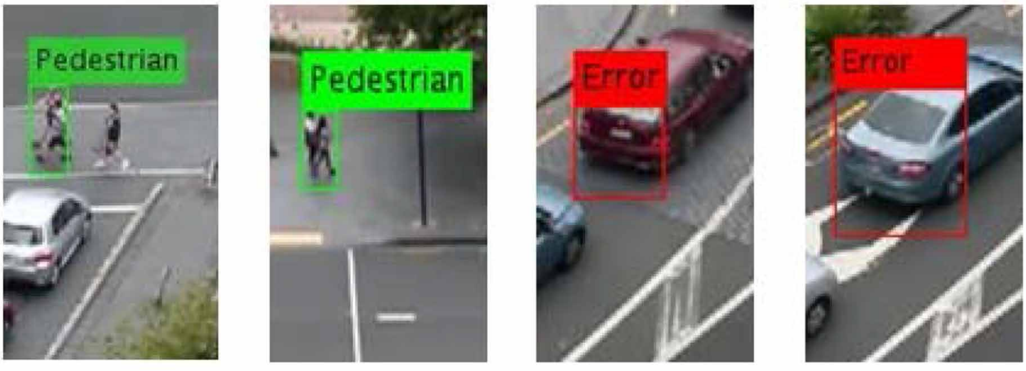


Figure 14. The results for pedestrian detection in real scenarios using HOG



For the ANN experiments, we utilised Matlab R2015b and Weka 3. For the *k*-Nearest Neighbour (*k*-NN) approach, the model performance will get affected by the *k* value. If there are too many neighbours in the *k*-nearest neighbour (i.e. *k* value too small), the resulting model will be *overfitted* and may learn from noisy samples and will not generalise well on future data. In contrast, when the *k* value is too large, the neighbour may include lots of points from other classes and the resulting model will produce suboptimal results. Therefore, the value of *k* should be chosen carefully, as in the experiments the *k* value is determined by the cross validation and should be less than the square root of the training set.

For the decision tree, there are three representative algorithms taken into consideration, named as: ID3, C4.5 and CART. The C4.5 algorithm addresses the shortcomings of ID3 that include avoiding the overfitting, handling the training dataset with the missing values and also improving the efficiency of the computation.

The learning algorithm for MLP used back-propagation (BP algorithm), the BP algorithm offers a gradient search technique which is to decrease the error function E as in Equation 1, where the y_p is the output, d_p is the desired output of the input pattern p:

$$E = \frac{1}{2}\sum_{p=1}^{n}\left(d_p - y_p\right)^2 \tag{1}$$

Moreover, it also presented the input samples as well as the corresponding desired output. For the principle of MLP, the network neurons calculate in the hidden layers until the output data shows each of the output values, after the data are presented into the input layer.

In the event recognition experiments, there were 780 samples of features as the training data for each behaviour, from six different videos. For cross-validation purposes, the dataset was split into 70% of training ratio (546 samples) and remaining both 15% of validation and testing data portions (117 samples each).

Histograms of Oriented Gradient (HOG) as one of the Feature Extraction Techniques (FET) is proposed into this paper and depicted in the previous chapters. In this paper, we will investigate the influence of HOG parameters on human behavior recognition. The comparisons between three feature extraction techniques will be explained in this section.

In the previous section, we pointed out that HOG feature adopts various parameters and cell sizes that directly affect the HOG descriptor. When the cell size is too small, it will increase the computational time of feature extraction. On the contrary, increasing the cell size could not include too much shape information which may affect result of the recognition.

Table 3 shows the results from multiple feature extraction techniques. As shown in this table, when the number of samples are the same, length of the feature vectors will affect the computational time. Moreover, LBP shows more efficient than HOG during the feature extraction.

Table 3. The results of different feature extraction techniques

Features	Samples	Cell Size	Feature Extraction Time	Feature Length
HOG	780	2-by-2	90.178s	30636
HOG	780	8-by-8	15.001s	1440
LBP	780	8-by-8	4.703s	3186
LBP	780	16-by-16	4.534s	708

For the classification testing in event recognition, we select two different videos for each of the five behaviours (Figure 15). Figure 16 illustrates the examples of event recognition for various human behaviors which are classified incorrectly. Figure 16(a) is based on HOG feature extraction. In the Skipping videos, there are three frames which are incorrectly classified into Running, Walking and Jump. In the Running test video, two frames which are incorrectly classified into Skipping. Figure 16(b) is based on LBP feature extraction. There are four different frames, which are classified incorrectly

Figure 15. The result of event recognition, which are correctly classified

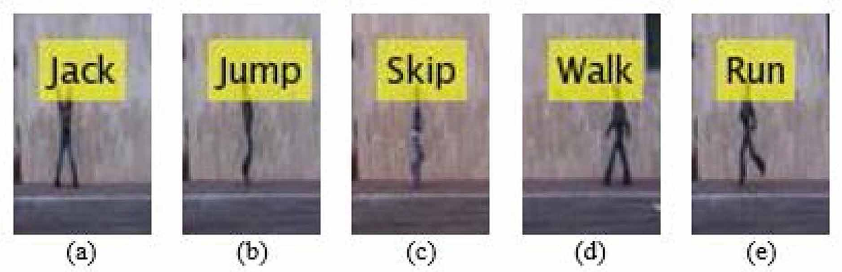
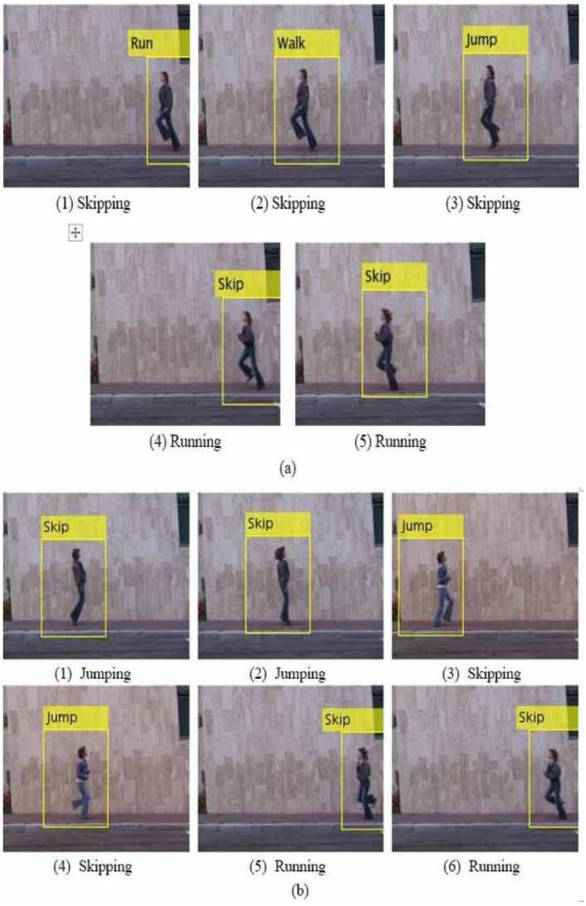


Figure 16. Incorrectly classified results of event recognition and classification based on (a) HOG feature extraction, (b) LBP feature extraction



into the Skipping class and also two frames are classified incorrectly into Jumping class. And Table 4 shows the results of classifications by using selected diverse classifiers for event recognition.

Table 4 shows results of the testing dataset with the precision of each human behavior. In the dataset, HOG feature achieves a 100% of precision in Jacking behavior recognition and the lowest precision is recognition rate of the human behavior Skipping which only has 94.2%. The LBP feature shows positive result in testing; the best precision is recognition rate of human behavior Jacking which achieves the precision at the rate of 96.2% and the lowest is 90.4%. If we use SILTP features, both Jumping and Running behaviors show 98.1% of precision rate and the lowest precision rate is human Skipping behavior.

Table 5 shows the best precision for detecting the human behaviours as jacking and jumping. However, the running and skipping behaviours have achieved lower precision in this experiment. The reason why we obtained these results is that the behaviours in these frames were similar to those associated with skipping activity. While LBP worked well (86.4%–94.9% classification accuracy) for pedestrian detection, due to spatial and temporal nature of the HOG method, HOG is well employed for both person detection (with 87.9%–95.6% classification accuracy) and human behaviour recognition (of 92.3%–100% classification accuracy).

CONCLUSION

In this paper, we demonstrated the feature extraction techniques such as Histograms of Oriented Gradient (HOG) and Local Binary Pattern (LBP) that could be adopted for classification purposes and modelling of human behaviour analysis.

By adjusting the cell size, the proposed empirical approaches are able to reduce the computational time of feature extraction. LBP feature which has the fastest extraction time is carried out in 4.5 seconds. Moreover, due to changes of various cell sizes which make the computational time greatly reduced. In feature extraction, HOG generates better results than those of LBP and SILTP which have greater potential in surveillance. The best precision of overall recognition for human behaviour achieves 97.7%, and both behaviours of Jacking and Jumping are recognized very well. Both features of HOG and LPB are relatively simple to implement in required low computational consuming so as to eliminate redundancy.

Understanding of feature extraction and visualisation of intermediated data processing represents a potential to inform future advancements. For human-intelligible visualisation of feature extraction, HOG produced better visual artefacts than LBP that has greater potential to inform and inspire

Table 4. The precision of different FET in the classification of testing dataset

	Jacking	Jumping	Skipping	Walking	Running	Total
HOG (8-by-8)	100%	98.1%	94.2%	98.1%	98.1%	97.7%
LBP (16-by-16)	96.2%	90.4%	94.2%	90.4%	92.3%	92.7%
SILTP	94.2%	98.1%	82.7%	94.2%	98.1%	93.5%

Table 5. Classification precisions using diverse classifiers

Classifier	Jack	Jump	Skip	Walk	Run
KNN	98.11%	100%	98.08%	98.08%	98.11%
Decision Tree	100%	98.11%	98.11%	96.15%	100%
MLP	98.11%	100%	94.55%	100%	92.31%

modelling decisions in pattern recognition and classification associated with human detection and behaviour recognition. Both HOG and LPB are relatively simple to implement requiring low computational resources to eliminate redundant data involved in video and image processing. However, there are still some limitations that should be improved in future. Because the testing videos only have single participant in each video, it is only suitable for private space. Only five of human behaviours were taken into account for recognition in this research project, in future we will use more natural human behaviours in our experiments.

Our future work includes:

1. We will work for human behaver analysis from multiple participants. In addition, more complex human behaviours such as fighting, robbery, etc. should be added in this project;
2. Since the testing surveillance videos were acquired from static cameras. In future, the high-speed camera could be taken into consideration in order to find more details of human behaviours, meanwhile more complex background should be taken into account. Furthermore, lighting conditions also could be improved which makes the method more operatable;
3. In future, other Feature Extraction Techniques (FET) should be taken into account to investigate human behaviour recognition. This is because the patterns of different behaviours need accurate and precise description;
4. The future work could be extended to develop further classifiers using deep learning and convolutional neural network.

REFERENCES

Bajpai, S., Jain, K., & Jain, N. (2011). Artificial Neural Networks. *International Journal of Soft Computing and Engineering*, *1*(NCAI2011), 27-31.

Burse, K., Manoria, M., & Kirar, V. (2011). Improved Backpropagation Algorithm to Avoid Local Minima in Multiplicative Neuron Model. In Information Technology and Mobile Communication (pp. 67–73).

Chen, X., & Zhang, C. (2006). An Interactive Semantic Video Mining and Retrieval Platform -- Application in Transportation Surveillance Video for Incident Detection. *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM '06)*, Hong Kong (pp. 129-138).

Cutler, R., & Davis, L. S. (2000). Robust Real-Time Periodic Motion Detection, Analysis, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 781–796. doi:10.1109/34.868681

Dai, K., Zhang, J., & Li, G. (2006). Video Mining: Concepts, Approaches and Applications. *Proceedings of the 2006 12th International Multi-Media Modelling Conference* (p. 4).

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886-893). doi:10.1109/CVPR.2005.177

Enzweiler, M., & Gavrila, D. M. (2008). Monocular Pedestrian Detection: Survey and Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(12), 2179–2195. doi:10.1109/TPAMI.2008.260 PMID:19834140

Geetha, P., & Narayanan, V. (2010). A Survey of Content-Based Video Retrieval. *Journal of Computer Science*, *4*(6), 734–734.

Intille, S. S., & Bobick, A. F. (2001). Recognizing Planned, Multiperson Action. *Computer Vision and Image Understanding*, *81*(3), 414–445. doi:10.1006/cviu.2000.0896

Ke, Y., Sukthankar, R., & Hebert, M. (2005). Efficient Visual Event Detection Using Volumetric Features. *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)* (Vol. 1, pp. 166-173).

Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., & Shafer, S. (2000). Multi-Camera Multi-Person Tracking for EasyLiving. *Proceedings of the IEEE International Workshop on Visual Surveillance* (pp. 3-10). doi:10.1109/VS.2000.856852

Laptev, I. (2005). On Space-Time Interest Points. *International Journal of Computer Vision*, *64*(2-3), 107–123. doi:10.1007/s11263-005-1838-7

Lu, J. (2016). Empirical Approaches for Human Behavour Analytics [Master's Thesis]. Auckland University of Technology, New Zealand.

Maryam, K., & Reza, K. M. (2012). An Analytical Framework for Event Mining in Video Data. *Artificial Intelligence Review*, *41*(3), 401–413.

Miller, J. F., & Khan, G. M. (2011). Where is the brain inside the brain? *Memetic Computing*, *3*(3), 217–228. doi:10.1007/s12293-011-0062-y

Mohan, A., Papageorgiou, C., & Poggio, T. (2001). Example-Based Object Detection in Images by Components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(4), 349–361. doi:10.1109/34.917571

Naphade, M. R., & Huang, T. S. (2002). Discovering recurrent events in video using unsupervised methods. *Proceedings of the International conference on Image Processing* (Vol. 2, p. 13). doi:10.1109/ICIP.2002.1039875

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *International Journal of Computer Vision*, *79*(3), 299–318. doi:10.1007/s11263-007-0122-4

Oliver, N. M., Rosario, B., & Pentland, A. P. (2000). A Bayesian Computer Vision System for Modeling Human Interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 831–843. doi:10.1109/34.868684

Popoola, O. P., & Wang, K. (2012). Video-Based Abnormal Human Behavior Recognition—A Review. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, *42*(6), 865–878. doi:10.1109/TSMCC.2011.2178594

Rui, Y., & Anandan, P. (2000). Segmenting Visual Actions Based on Spatio-Temporal Motion Patterns. *Proceedings of the IEEE Conference on Computer Vision* (pp. 111-118). doi:10.1109/CVPR.2000.855807

Siebel, N., & Maybank, S. (2002). Fusion of Multiple Tracking Algorithms for Robust Prople Tracking. *Proceedings of the European Conference on Computer Vision* (pp. 373-387).

Sondak, N. E., & Sondak, V. K. (1989). Neural networks and artificial intelligence. *Proceedings of the twentieth SIGCSE Technical Symposium on Computer Science Education* (pp. 241-245). doi:10.1145/65293.71221

Su, Z., Zhang, H., Li, S., & Ma, S. (2003). Relevance Feedback in Content-Based Image Retrieval: Bayesian Framework, Feature Subspaces, and Progressive Learning. *IEEE Transactions on Image Processing*, *12*(8), 924–937. doi:10.1109/TIP.2003.815254 PMID:18237966

Tran, S. D., & Davis, L. S. (2008). *Event Modeling and Recognition Using Markov Logic Networks. Computer Vision–ECCV* (pp. 610–623). Berlin, Heidelberg: Springer.

Tziakos, I., Cavallaro, A., & Xu, L. Q. (2010). Event Monitoring Via Local Motion Abnormality Detection in Non-linear Subspace. *Neurocomputing*, *73*(10), 1881–1891. doi:10.1016/j.neucom.2009.10.028

Valera, M., & Velastin, S. A. (2005). Intelligent Distributed Surveillance Systems: A Review. *IEE Proceedings. Vision Image and Signal Processing*, *152*(2), 192–204. doi:10.1049/ip-vis:20041147

Velipasalar, S., Brown, L., & Hampapur, A. (2006). Specifying, Interpreting and Detecting High-level, Spatio-Temporal Composite Events in Single and Multi-Camera Systems. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, Washington, DC, USA (p. 110). doi:10.1109/CVPRW.2006.197

Wang, L., & He, D.-C. (1990). Texture Classification Using Texture Spectrum. *Pattern Recognition*, *23*(8), 905–910. doi:10.1016/0031-3203(90)90135-8

Westermann, U., & Jain, R. (2007). Toward a Common Event Model for Multimedia Applications. *IEEE MultiMedia*, *14*(1), 19–29. doi:10.1109/MMUL.2007.23

Xie, L., Sundaram, H., & Campbell, M. (2008). Event Mining in Multimedia Streams. *Proceedings of the IEEE*, *96*(4), 623–647. doi:10.1109/JPROC.2008.916362

Xie, L., & Yan, R. (2009). Extracting Semantics from Multimedia Content: Challenges and Solutions. In *Multimedia Content Analysis, Signals and Communication Technology*.

Zelnik-Manor, L., & Irani, M. (2001). Event-Based Analysis of Video. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 123-130).