Retrieval of Distinctive Regions of Interest from Video Surveillance Footage: A Real Use Case Study

C.A. Mitrea[1,2], T. Piatrik[2], B. Ionescu[1], M. Neville[3]

[1]Image Processing and Analysis Lab, UPB, Romania, *catalin.mitrea@uti.ro*, *bionescu@imag.pub.ro*
[2]Multimedia and Vision Group, QMUL, UK, *t.piatrik@qmul.ac.uk*
[3]Central Forensic Image Team, MET, UK, *mick.neville@met.police.uk*

Keywords: Distinctive regions of interest, content-based retrieval, video surveillance, real-world data.

Abstract

The article addresses the issue of retrieving distinctive regions of interest or patterns (DROP) in video surveillance datasets. DROP may include logos, tattoos, color regions or any other distinctive features that appear recorded on video. These data come in particular with specific difficulties such as low image quality, multiple image perspectives, variable lighting conditions and lack of enough training samples. This task is a real need functionality as the challenges are derived from practice of police forces. We present our preliminary results on tackling such scenario from Scotland Yard, dealing with the constraints of a real world use case. The proposed method is based on two approaches: employment of a dense SIFT-based descriptor (Pyramidal Histogram of Visual Words), and use of image segmentation (Mean-Shift) with feature extraction on each segment computed. Tested on real data we achieve very promising results that we believe will contribute further to the ground development of advanced methods to be applied and tested in real forensics investigations.

1 Introduction

High expansion of urban population and recent world events have mobilized major industry players to redesign physical security concepts. Consistent efforts are channeled to reinforce authority's actions on coping with the main drawbacks and issues generated by the growing rate of urbanization. The latter, combined with the continuous threats to public and infrastructure safety, are contributing to the exponential increase of the number of video surveillance cameras, e.g., it is estimated that there are more than 4 millions CCTV cameras deployed in UK[1]. Some of the major drawbacks of existing video surveillance systems are the absence of efficient data processing and the lack of a distinctive regions of interest or patterns (DROP) retrieval architecture. For instance, once a DROP was identified in a frame by an operator, the existing approaches provide poor identification and tracking capabilities in previous and current multiple-source recordings. This task is actually accomplished manually by human operators, many times being highly time consuming and inefficient.

Retrieval of DROP instances from video surveillance footage involves processing of challenging low quality images for which established methods of feature detection, extraction and matching perform less well than in high-definition images [13]. The quality as well as the viewing perspective varies considerably for each security camera, this inconsistency usually roughening algorithms performance. Images are often low resolution with poor color clarity and have little discriminative or representative texture definition. The footage is often recorded on fixed or moving cameras (e.g., PTZ - Pan Tilt Zoom based) to provide maximum coverage of an area with a single camera. Objects will therefore move in and out of view regularly within a sequence of frames while the movements can denote dramatic changes of focus and rapid zooming, worsening further the investigations. On these premises, a high performing automated video surveillance becomes highly necessary.

In this work we investigate two approaches that use confirmed techniques from the literature. The first is based on keypoints features (i.e., Pyramidal Histogram of Visual Words - PHOW), and the second, is driven by image segmentation (via Mean-Shift) and feature extraction on each object generated. The novelty of our approach relies on means and practices on which the algorithms are adapted and parameters are calibrated and optimized to cope with a real world task. We present our preliminary results in this direction.

The rest of the paper is structured as following. Section 2 investigates existing techniques and advances in literature related to our task and situates our contribution. Section 3 presents the architecture of the proposed approach. Experimental validation and discussion of the results is presented in Section 4 while Section 5 concludes the paper.

2 Related Work

Automated analysis of security video sequences is still an open issue that demands adaptation of existing technology to cope with increasing challenges, such as the ones mentioned in the previous section, e.g., low image quality, lack of training samples, etc. Some of the directions and research topics in

[1]http://www.securitynewsdesk.com/bsia-attempts-to-clarify-question-of-how-many-cctv-cameras-in-the-uk/

video surveillance systems include: detection, categorization and tracking of objects of interest in video (e.g., people, vehicles, abandoned baggages) [8, 23, 18], recognition of their activities (e.g., event and behavior analysis) [22], and efficient technologies to process large amounts of data (Big Data) [7]. Generally all these tasks more or less follow a basic framework. The object is detected, some features are computed, then based on a "decisioner" (a classifier), the object is labeled and further placed to a higher level for event and behavior analysis (e.g., using ontology techniques or other higher knowledge representations technologies) [19].

Most of the work related to our task and addressed in the literature report contributions at the level of efficient video content description while leveraging the discriminative power. For pattern recognition tasks, most of methods revolves around low-level feature retrieval, many image feature extraction algorithms being proposed, e.g., color [14], texture [21], shape [1], or the popular feature point descriptors combined with Bag-of-Words [13, 17].

Generally all methods include edge, corner, blob and region detectors and assume that the features detected from sample images should remain unchanged under geometric transformations, which would enable proper matching of images taken at different views and points in time. Most of these algorithms' success is empowered, at some degree, by robustness to rotation, change of scale, illumination shifting or even signal perturbations. Some other approaches are contributing to the level of effective decision making. At this point, the reasoning is usually powered by classifiers or distance metrics [11]. Some of them are investigating methods of automatic pre-processing and refinement of input data in order to leverage classifier accuracy [20]. Other are focusing on parameters tuning in order to cope with effective training and data noises [3].

One drawback of the current state-of-the-art "decisioners" is the low generalization power when insufficient training samples are available. This is a particular issue in video surveillance, as most of the time, the available instances of the target object to be searched for is limited, i.e., only few seconds of footage, or only few images are available to formulate the query.

As a general observation, usually object detection and recognition methods in video surveillance are relying on motion information. In real world scenarios the video cameras are usually PTZ based, therefore common motion driven techniques [6] (e.g., background substraction) might be less effective. As we need to find DROP instances on entire dataset starting from just one single sample, traditional decision making powered by classifiers is less suitable.

To cope with aforementioned limitations, the proposed method exploits the benefits of key-points features and image segmentation methods, which at some degree are robust to motion, while the decision is achieved using distance matching of the feature vectors. The entire experimentation is carried out on Scotland Yard footage and the task is a real use case of police practices. Findings and output of this research are contributed to the understanding of the constraints and issues that are particular to real-world video surveillance datasets.

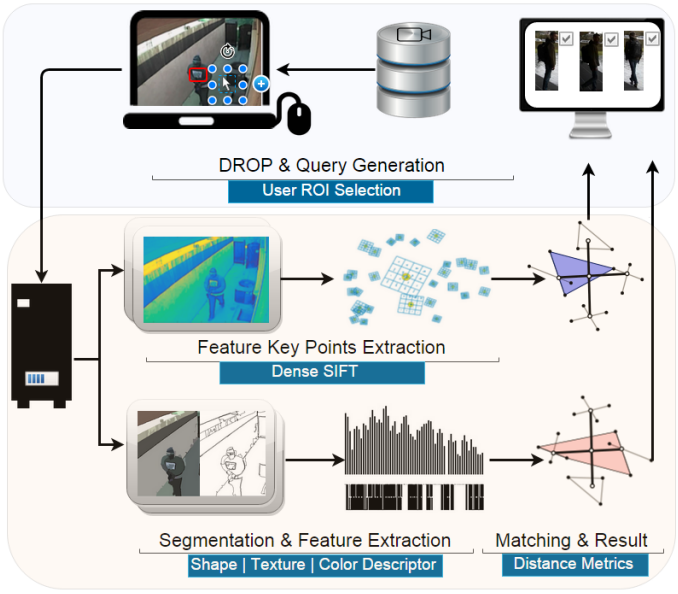

Figure 1. Proposed system architecture.

3 System Architecture

The system built to conduct the experiments is composed mainly of two layers (see Figure 1).

First layer deals with sample selection and query generation. At this layer the police officer selects (by fitting a rectangle on the DROP area) the distinctive region of interest to be searched for, many times having at his disposal only one image available.

Second layer deals with video data processing using two approaches. The first approach (see Section 3.1) is based on a highly dense SIFT [15] algorithm (PHOW - Pyramidal Histogram of Visual Words) [2] while the second approach (see Section 3.2) is powered by a Mean-Shift segmentation [5] complemented by several feature extraction techniques.

These two approaches were selected due to their close appropriateness to our underlying scenario. Basically the SIFT descriptor is invariant to scaling, rotations and translations in the image domain and robust to moderate perspective transformations (e.g., image warping) and even to some degree of illumination variations. Nevertheless it tends to fail on providing a more complete description of the image content (e.g., shape or color scene appearance). Therefore, the second approach is driven by Mean-Shift segmentation on which a more complete description of scene content is employed, i.e., shape-color-texture analysis. The Mean-Shift algorithm is a simple but effective method for estimating the density gradient. One tremendous advantage for selecting it in the current work is the nonparametric nature as it does not presume that the segments resulted are derived from a specific probability distribution. The implementations are based on VLFeat[2] computer vision library.

[2]http://www.vlfeat.org/

3.1 Key-point feature-based matching

In the current work we have adapted a dense SIFT based descriptor, i.e., Pyramidal Histogram of Visual Words (PHOW). Main steps of this approached are: (i) Extract PHOW features [2]; (ii) Match between sample PHOW features and PHOW extracted from each video frame (algorithm suggested in [15]); (iii) Refine further matches by using RANSAC (RANdom Sample and Consensus) [9] algorithm with homography model [4]; and finally (iv) Based on a minimum distance, matches are grouped and the results are returned to the operator.

We have employed the PHOW descriptor as the region of interest can be very small. The PHOW features are a variant of dense SIFT descriptors, extracted at multiple scales. This typically generates a very large number of features. For example on a 50×60 pixels DROP sample, normal SIFT extracts ca. 25 key points while the PHOW technique computes ca. 825 key points (therefore 800 more key points which should increase matching results). Secondly, we have adapted the RANSAC with the homography model as it reduces the matches that are too ambiguous. RANSAC is a learning technique which estimates parameters based on a random sampling model of observed data. Given the matching points whose elements contain both inliers and outliers, the algorithm uses the voting scheme to find the optimal fitting matching points which further is optimized using homographies correlations between sample and frame points, therefore filtering out more false matches. We will refer further to this approach using the acronym PHOW.

3.2 Segmentation/feature-based matching

The second method employs a segmentation and feature extraction on each segment computed. The main processing steps of this approach consists of: (i) Each frame is segmented using Mean-Shift algorithm [5]; (ii) On each segment a feature descriptor is computed; (iii) A matching is calculated between sample and each segment; and finally (iv) Results are returned to the operator.

Mean-Shift segmentation is a procedure for locating the maxima of a density function given discrete data sampled of that function. Main idea for this method is to extract objects and features which further are matched with the DROP sample. For matching we are using a ChiSquare distance metrics [11].

The following video descriptors which denoted good performance in other related tasks [18] were employed for feature extraction:

- CSD (Color Structures Descriptor) [16] is based on color structure histogram (a generalization of the color histograms) to encode information about the spatial structure of colors in an image as well as their frequency of occurrence. The resulting vector has 32 features;

- HOG (Histogram of Oriented Gradients) [12] shape based descriptor, the algorithm counts occurrences of gradient orientation in localized portions of an image, computed on a dense grid of uniformly spaced cells which are further composed and combined to generate the final vector of 81 features;

Figure 2. Example of distinctive region of interest - DROP selection conducted by police officers. This sample is used to query and retrieve similar DROP instance from the dataset.

- LBP (Local Binary Patterns) [10] texture based descriptor, represents a particular case of the texture spectrum model, based of a simple texture operator which labels the pixels on an image cell by thresholding the neighborhood of each pixel and outputting the result as a binary number. The resulting vector has 256 features.

We will refer further to this approach using the acronym SEGM.

4 Experimental Results

4.1 Dataset

The proposed system was experimented on real world data consisting of footage acquired from CCTV cameras provided by Scotland Yard. In London it is estimated there is one CCTV camera for every 14 people, which would generate a number of more than 400,000 units deployed in the city. Such vast number generates an impressive volume of video data. Based on common CCTV recording parameters (704 x 576 pixels, ca. 10 frames per second, H.264 codec and 12 hours of motion activity per day) this setup will produce in 24 hours around 2.5 GB of footage per recording channel, and overall, a total of 1 million of GB each day (ca. 30,000 TB each month).

In this paper we present our preliminary results obtained on a sample of these data consisting of three recordings, summing 115 seconds of video, generating ca. 1,400 frames (see samples in Figure 3). The target DROP is present or visible in ca. 350 images (size ranging from ca. 75 x 65 to 25 x 20 pixels).

4.2 Real Use Case

The used data contain a real world scenario of the Metropolitan Investigation Police that addresses identification of suspects. The footage was acquired during the 2011 riots[3] that took place in London. In Figure 2 is depicted an example of DROP selection used for tracking and identification performed by police operators. Here the distinctive "back logo" of the perpetrator (which is vandalizing/breaking police car window, see Figure 3-1) is used for person identification in subsequent images.

[3] http://www.theguardian.com/uk/2011/aug/08/london-riots-escalate-police-battle

Starting point of the analysis is the identification of suitable images related to the scene of the crime. These images are then used to identify relevant DROP instances in the entire database.

In this particular example it becomes clear that direct face identification would fail. Thus, other distinctive regions or patterns related to the criminal are used for further search. Therefore, the first image illustrating the criminal act is used as starting point for the investigation. The goal is to retrieve all the instances of humans that have that pattern and which finally could offer a clear (face) identification of the suspect.

Following section deals with parameters setup for both proposed approaches.

4.3 Parameters tuning

For the first approach, a DROP key-point feature is considered matched to a key-point feature extracted from the frame only if the (squared Euclidean) distance between them, multiplied by a threshold, is not greater than the distance to the rest of the key-points features computed from the frame. The threshold value in this work is selected with a value of 1.2. RANSAC is a "best effort" iteration process, for this work we selected 25 iterations (chosen on empirical evidence) for determining optimal parameters to fit the model.

For the second approach, segmentation process performance is based on three kernel parameters. Spatial and color bandwidth were set to a value of 9, respectively 5. These two parameters control the convergence of the density segmentation function. Other parameter is the minimum (size) of segmented region (set to a value of 100 pixels) that controls the number of segments generated (smaller areas are merged to the neighboring ones). This setup will generate on average around 250 segmented regions (homogeneous tiles) per frame.

4.4 Performance evaluation

To evaluate properly the system it was necessary to build a ground-truth (GT) by manually labeling all the frames (draw coordinates) in which the DROP is present.

To our task (i.e., retrieval of distinctive region of interest) it is more important to retrieve all the existing instances with the risk on including false matches rather than obtaining few false matches but missing some of the existing instances. To assess performance we use the standard F-score measure where we consider precision (accounts for the number of false positives) as weighting higher than recall (accounts for the non-detections), thus:

$$F_\beta = (1+\beta^2)\frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad . \qquad (1)$$

where *Precision* is computed as TP/(TP+FP), *Recall* is TP/(TP+FN) and TP are the True Positives (correct detections, e.g., the DROP instance is consistent with GT), FP represents the False Positives and FN are the False Negatives. As the emphasis is on precision, in particular $\beta = 0.5$.

In Table 1 are depicted the results in terms of F-score (in percents) for the DROP scenario, e.g., searching for the "back logo" (see Figure 2). Best performance is obtain using the PHOW approach (i.e., 57.15). Second best result is obtained with the SEGM approach (i.e., 43.39) using CSD descriptor while de lowest score is obtain by the SIFTs (30.67).

Table 1. System performance results.

	Key-point based matching		Segmentation/feature based matching		
Descriptor	SIFT	PHOW	CSD	LBP	HOG
F-Score	30.67	57.15	43.39	37.42	41.64

As the DROP sample gets very small (ca. 50 x 50 pixels) the normal SIFT descriptor extracts on average few tens of key-points which are simply not enough in the matching process. This is not the case of PHOW descriptor which is generating consistently more key-points (few hundreds) which further increases the chances of matching.

Overall the PHOW approach denotes better performance compared to SEGM approach. Anyway the SEGM approach denotes also suitable performance as all three selected descriptors (CSD, LBP and HOG) are powerful feature extractors which have been used with success in many pattern recognition and extraction tasks [18, 12, 10, 16].

In Figure 3 are shown some sample images of the footage. Green frame rectangle shows positive DROP retrieved (1,2,3) while with red color are depicted the instances what were missed by the system (4,5,6). Reasons of misdetection is the low size of the DROP which is not providing enough discriminative information. In Figure 3-1 the suspect is vandalizing the police car while in Figure 3-6 the suspect throws a step-ladder to police officers.

4.5 Strengths and limitations

Each method has its own advantages and limitations. Both approaches are somehow robust to some image transformations (as scaling and translation). From the experiment above, they are able to retrieve the DROP to a specific minimum scale threshold (from measurements - around 35 x 30 pixels, see Image 3-3). Below that threshold both methods are denoting a decreasing of performance.

Compared to second method, PHOW approach is denoting better performance to slight lighting variations and translations. Anyway the method fails if the DROP texture is faded or too blurred. Latter does not apply to SEGM approach as the Mean-Shift algorithm is able to extract the object while supported by the CSD color descriptor.

SEGM approach is more suitable for content-based retrieval, clustering and indexing as on each frame computed, if required, extracted objects can be clustered based on their features. This provides some advantages. For example if all the objects are indexed, on another query, the process of DROP instances retrieval should be faster as there is no need for image segmentation (e.g., to re-run the entire processing chain), instead only to determine on which cluster best fits the new DROP. Also, the method provides more scene "insights" or

Figure 3. Examples of True Positives (TP - green frame) and False Negatives (FN - red frame) detections. Note the different DROP sizes and overlapping with other objects (2,3).

scene appearance (e.g., shapes or objects colors) which makes it suitable and easy to use with upper, high level of knowledge processing and representation (as ontology). One drawback of the SEGM approach is the parameters' selection. For example, if minimum region size parameter is setup higher than 200 pixels, the algorithm might fail to detect DROP instances that are very small. Setting the value too low (less that 20 pixels) will generate many image segments (thousands) which increases computation complexity.

Automated video surveillance and monitoring is generating many issues such as dealing with occlusion and object interaction in high density scenes (shadows, occlusion and cluttering). Also tracking across multiple overlapping and non-overlapping field cameras are some of the main challenges in DROP retrieval. Analysis is further reduced by varying weather conditions where the changes in light, presence of rain, snow, mist or fog, direct sunlight and shadows can all affect the clarity of an image.

5 Conclusion

In current work we have investigated and compared (obtaining acceptable performance in terms of F-Score) two different approaches for coping with retrieval of distinctive regions of interest or patterns (DROP) from video surveillance footage.

Figure 4. Samples of segmented images (1,3) and key points matches (2,4).

For real-world datasets "academic" state-of-the art algorithms needs to be adapted to new challenges: how to deal with low quality/noisy data sets and how to "learn" starting from few (or one as in our case) training samples. For example it has been shown that if DROP sample size is small, standard SIFT descriptor shows difficulties on retrieving enough key points to ensure matching success.

Further work needs to address some other specific issues like the impact of image quality on the matching process, knowledge transfer (inheritance) for later usage and computational complexity (method's power to provide "real-time" responses to queries). For latter, one approach is to adopt Big Data technologies to cope with large scale of datasets and manipulation constraints (e.g. Hadoop-based video processing). Further investigations are required on: (i) fusion of both approaches to make use of combined benefits; (ii) artificial data generation, e.g., controlled image warping (non planar rotation) in order to induce information on DROP geometrical transformations; (iii) use of relevance feedback (RF). Once a new DROP instance is correctly identified, new information can be embedded to the reasoner to enhance performance; (iv) while the system evolves and retrieves more "good" DROP instances, a dictionary can be created or further some machine learning techniques can be applied. Latter has the main advantage of inducing into the system reasoning complex new relations and knowledge of the DROP instances.

Acknowledgements

We are thankful to Metropolitan Police Service for their valuable guidelines and providing the data. This work is partially supported by the LASIE[4] project (grant agreement nr. 607480)

[4]http://www.lasie-project.eu/

and also by ExcelDOC POSDRU/159/1.5/S/132397, PN-II-IN-DPST-28DPST/30.08.2013 and PN-II-PT-PCCA-290/2014.

References

[1] X. Bai, C. Rao, and X. Wang. Shape vocabulary: A robust and efficient shape representation for shape matching. *IEEE Transactions on Image Processing*, pages 3935 – 3949, 2014.

[2] A. Bosch, A. Zisserman, and X. Munoz. Image classifcation using random forests and ferns. *ICCV Processing*, 2007.

[3] O. Chapell, V. Sindhwani, and S.S. Keerthi. Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, pages 203–233, 2008.

[4] O. Chum, T. Pajdla, and P. Sturm. The geometric error for homographies. *Computer Vision and Image Understanding*, pages 86–102, 2005.

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 603–619, 2002.

[6] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1337–1342, 2003.

[7] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Search Storage, Google, Inc.*, 2004.

[8] C.R. del Blanco, F. Jaureguizar, and N. Gacia. An efficient multiple object detection and tracking framework for automatic counting and video surveillance applications. *IEEE Transactions on Consumer Electronics*, 58:857–862, 2012.

[9] A. Hast, J. Nysjo, and A. Marchetti. Optimal ransac towards a repeatable algorithm for finding the optimal set. *Journal of WSCG*, pages 21–30, 2004.

[10] D.C. He and L. Wang. Texture unit, texture spectrum, and texture analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 28:509–512, 1990.

[11] R. Hixon and D.M. Gruenbacher. Evaluation of the fisher discriminant and chi-square distance metric in network intrusion detection. *Annual Technical and Leadership Workshop*, pages 119–124, 2004.

[12] L. Hu, W. Liu, B. Li, and W. Xing. Robust motion detection using histogram of oriented gradients for illumination variations. *2nd International Conference on Industrial Mechatronics and Automation (ICIMA)*, 2:443 – 447, 2010.

[13] G. Khan, N.Y. McCane, and B. Wyvill. Sift and surf performance evaluation against various image deformations on benchmark dataset. *International Conference on Digital Image Computing Techniques and Applications (DICTA)*, pages 501 –506, 2011.

[14] R. Khan, J. Weijer, F. Khan, D. Muselet, C. Ducottet, and C. Barat. Discriminative color descriptors. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2866–2873, 2009.

[15] D. G. Lowe. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] B. S. Manjunath, J.R. Ohm, V.V. Vasudevan, and A. Yamada. Color and texture descriptor. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 703–715, 2001.

[17] I. Mironica, I. Duta, B. Ionescu, and N. Sebe. Beyond bag-of-words: Fast video classification with fisher kernel vector of locally aggregated descriptors. *IEEE International Conference on Multimedia and Expo - ICME*, pages 501 –506, 2015.

[18] C.A. Mitrea, I. Mironica, B. Ionescu, and R. Dogaru. Multiple instance-based object retrieval in video surveillance: Dataset and evaluation. *IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 171–178, 2014.

[19] J.C. SanMiguel, J.M. Martinez, and A. Garcia. An ontology for event detection and its application in surveillance video. *Advanced Video and Signal Based Surveillance*, pages 220–225, 2009.

[20] T.R. Tavares, G.G. Cabral, and S.S Mattos. Preprocessing unbalanced data using weighted support vector machines for prediction of heart disease in children. *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013.

[21] N. Vatani, M. De Deuge, B. Douillard, and S.B. Williams. An analysis of monochrome conversions and normalizations on the local binary patterns texture descriptors. *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 831 – 837, 2013.

[22] B. Yogameena and K.S. Priya. Synoptic video based human crowd behaviour analysis for forensic video surveillance. *Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, pages 1–6, 2015.

[23] S. Zhang, C. Wang, S. Chan, X. Wei, and C. Ho. New object detection, tracking, and recognition approaches for video surveillance over camera network. *IEEE Sensors Journal*, 15, 2015.