Deeply Self-Taught Multi-View Video Analytics Machine for Situation Awareness

Ming Shao, and Yun Fu

Organization: Northeastern University

Address: 360 Huntington Ave., Boston MA 02115, USA

Telephone: +1 (857)288-9393

Email: {mingshao, yunfu} @ece.neu.edu

Focus area: Situation awareness

Total words: 2950

Deeply Self-Taught Multi-View Video Analytics Machine for Situation Awareness

Ming Shao, *Student Member, IEEE,* and Yun Fu, *Senior Member, IEEE*

***Abstract*—Situation awareness involves being aware of what is happening in the vicinity, in order to understand how information, events, and one's own actions will impact goals and objectives. In this paper, we target at video analytics aspect of situation awareness and propose a new model for anomaly detection in a huge amount of image/video data. The proposed model is called "Deeply sElf-taught multi-view Video Analytics machiNe (DEVAN)", which is able to handle data from either manned or unmanned air craft, and report anomaly detection results in a fast manner by analyzing signals from multiple channels. The proposed model is composed of three components. First, multi-view learning module is able to fuse features from different sensors, and ensure to discover different levels of anomaly that might be ignored by a single channel. Second, we propose a deep robust feature extractor that refines feature from coarse to fine along with the multi-view learning. Third, we design a self-taught learning procedure to explore anomaly only from the data itself and its underlying distribution, without additional hand labeling work. In the last, we also clarify the datasets and experimental settings for system evaluations.**

***Index Terms*—Sitation awareness, anomaly detection, deep learning, multi-view learning**

I. Motivation

Situation awareness (SA) means appreciating all you need to know about what is going on when the full scope of your task [1]. More specifically and in the context of complex operational environments, SA is concerned with the person's knowledge of particular task-related events and phenomena. In air force missions, this is especially important for the fighter pilot who should be aware of the threats and intentions of enemy forces as well as the status of his/her own aircraft [2]. It is even more important for intelligent agency to analyze massive data collected from drone footage in an efficient way [3]. For example, in 2011, unmanned aerial vehicles (UAV) collected 327,384 hours of video from surveillance cam-equipped UAVs, and it might need almost the same working hours of human to do exhaustive anomaly event search. Such event usually hides behind messy background, massive moving objects, and pedestrians. In addition, the data source is diverse and video quality is not guaranteed. In recent Boston bombing event[1], the suspects are finally targeted through the videos from spectators after hours of investigations by technicians, although the scale data is not large. Such brute-force man power based search is very impractical for drone footage, which scales largely in both spatial and temporal domain.

Although unmanned aircraft vehicle is increasingly popular for years in military (See Figure 1), the relative video analytics techniques still fall behind. It has been said that "Military Is Awash in Data From Drones". which is under the background that there were already millions of investment on computer systems for drone footage analysis, and 4000 airman employed for such work in 2010.

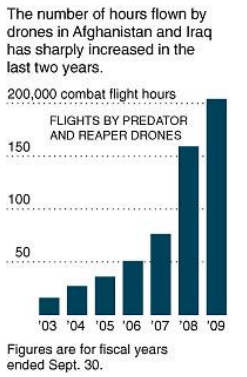


Fig. 1. Running hours of drone is increasing in last a few years [3].

In fact, video analytics for anomaly events has been widely discussed in the computer vision community for years [4], [5], [6], [7], [8], [9], [10], [11], which is urged by the strong demands from government and public safety. However, existing techniques suffer from the following aspects. First, most of these methods rely on supervised learning models and labeled data. Second, the low-level visual descriptors are fragile and affected by noises and corruptions of data due to uncertainty. Third, they only concentrate on single view/source signal based anomaly events. Therefore, a direct application of existing techniques can not satisfy the requirements of large-scale, robust, and multi-view anomaly detection raised by drone footage video analytics.

II. Introduction

As a fundamental problem, anomaly detection motivates diverse research areas and application domains. It is approaching issues in certain application domains through techniques specifically developed, as well as those are more generic. An excellent survey can be found in [12], where a structured and comprehensive overview of the research on anomaly detection is provided. In addition, they categorized state-of-the-art methods based on the underlying approach adopted by each of them.

Video analytics drew great attentions ever since 2000, when the Advanced Research and Development Activity (ARDA) started sponsoring detection, recognition, and understanding of moving object events, which focused on news broadcast video, meeting/conference video, unmanned aerial vehicle (UAV) motion imagery and ground reconnaissance video, and surveillance video. The Defense Advanced Research Projec-

M. Shao is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA. E-mail: mingshao@ece.neu.edu.

Y. Fu is with the Department of Electrical and Computer Engineering and College of Computer Information Science, Northeastern University, Boston, MA 02115 USA. E-mail: yunfu@ece.neu.edu.

[1]https://en.wikipedia.org/wiki/Boston_Marathon_bombing

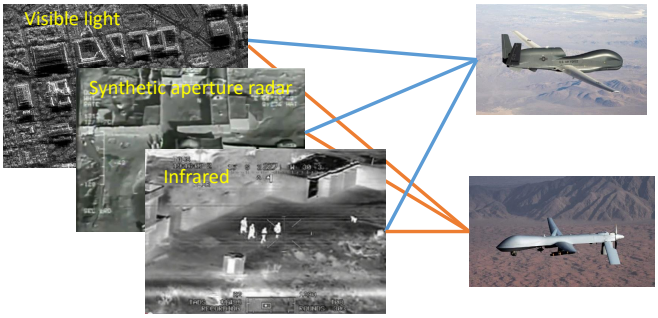


Fig. 2. Illustrations of multi-view data from UAV, e.g., multi-aircraft (right), multi-sensor (left).

tion Agency (DARPA) has also supported several large research projects involving visual surveillance and related topics. Projects include Visual Surveillance and Monitoring (VSAM, 1997) and Human Identification at a Distance (HID, 2000). The most recent project, Video and Image Retrieval Analysis Tool (VIRAT 2008) aims to develop and demonstrate a system for UAV video data exploitation, enable quick response or alert of an event or retrieve video content from achieves.

Video analytics has also been appealing to commercial systems. Readers could refer to a summary of commercial systems in an good survey [13], where they list advertised capabilities for human behavior recognition. However, among many existing systems, it is still difficult to quantitatively measure the performance of each of them since tasks handled are treated differently by agencies with different aims or focuses. Therefore, many efforts have been devoted to standard evaluation frameworks, i.e., methodologies to quantify and qualify performance [14], [15], [16], [17], [18], [19].

Multi-view anomaly detection. Multi-view data analysis has caught a great deal of attention in the recent years [20], [21], [22]. However, multi-view anomaly detection has not been discussed before, which however is common in many real-world scenarios. Anomaly can be detected by different sensors of an UAV, or sensors from different UAVs (Figure 2). Such anomaly may be caused by: (1) unstable factors caught by one sensor, (2) emergency caught different sensors. To the best of our knowledge, this is the first time when different levels of anomaly are discussed under a multi-view learning framework.

Self-taught low-rank anomaly representation. Modeling anomaly events is not an easy task as there are few labels available for supervised learning algorithms due to the diversity of anomaly. In this paper, we propose a new anomaly modeling methods based on representation learning. Specifically, for each view, we decompose the anomaly features into a low-rank and a sparse components to capture the normal feature space, and anomaly feature vectors, respectively. By such decomposition, we could learn efficient representations from data itself, without knowing any label information. Therefore, all discriminative knowledge is learned from data itself and corresponding underlying distributions.

Deeply refined robust feature. Recently, deep structure and its modeling have attracted substantial research attentions from computer vision, and machine learning communities due to its appealing performance on many challenging performance, e.g., face recognition in the wild [23] and 1000-classes objects classification challenge [24]. The multi-layer deep structure could build features from coarse to fine, and the denoising mechanism is able to rule out outliers and local perturbations of data. In this paper, we propose to integrate such merits with our self-taught anomaly representations, and therefore are able to address the data noise in a progressive way.

III. Challenges in Proposed Work

Uncertainty of multi-view anomaly detection. Anomaly detection in single view has long been debated; however, fewer works have mentioned multi-view anomaly detection. The key issue is the target may be recognized as anomaly in one view, but as normal event in another view. In addition, anomaly events may have different semantics in different views, which is reflected by the structure of feature space. Therefore, handling different types/levels of anomaly in one framework is very challenging.

Anomaly detection by unlabeled data. Supervised anomaly detection relies on the well-defined anomaly events and labels on the training set. However, nowadays, "data explosion" keeps challenging the state-of-the-art algorithms in computer vision. Manually labeling huge dataset for training is already impractical, especially for anomaly detection, because the definition of "anomaly" heavily depends on domain knowledge. Therefore, designing an algorithm that is able to explore the patterns of anomaly automatically by feature space structure and detected feature portions makes great sense.

Large-scale video analytics. Although nowadays drone footage video data have already flooded the intelligence agency, the recent research achievements on "big data" have demonstrated the necessity of "large-scale" on challenging vision problems. In our system, we will handle large-scale problems in three aspects: (1) low-rank anomaly representation; (2) anomaly categorization; (3) deeply refined feature, whose computational complexity are highly related to the number of samples in the dataset.

IV. Technical Proposal

A. Framework Overview

The system overview is shown in Figure 3, including four functional parts: (1) low-level feature extractor, (2) self-taught low-rank representation, (3) multi-view anomaly detection, and (4) deep robust feature refinement.

B. Low-Level Feature Extractor

Low-level features for events are sparsely distributed in frames without clear semantics. However, they are highly related to the motion features, and therefore, we could locate important low-level features by detecting motions first. Second, we will refine the tracking results by priors such as the shape, color, and texture of human/object. Third, we will extract low-level visual descriptors only on the interested area.

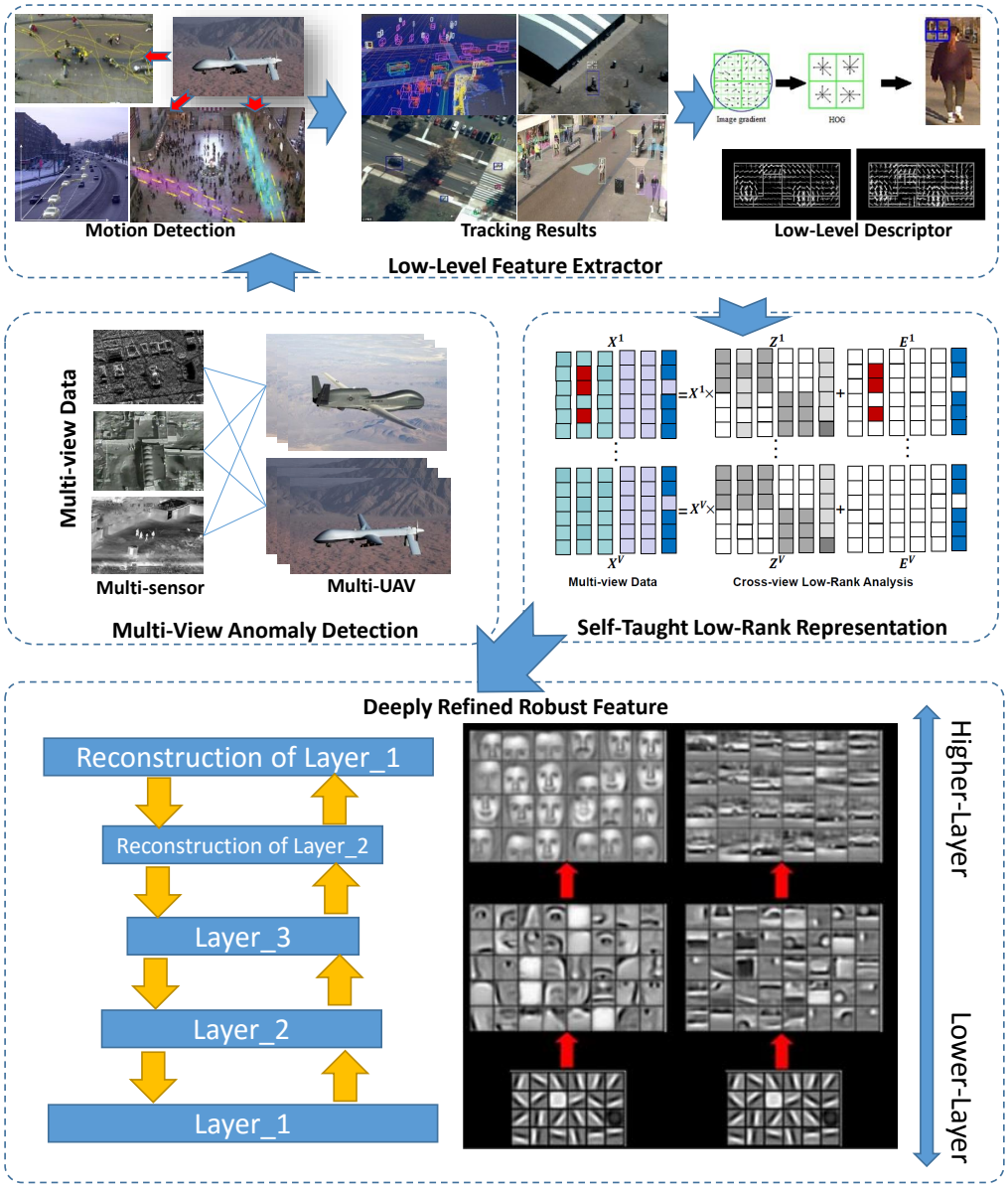


Fig. 3. Framework illustration of our method including four components: (1) low-level feature extractor, (2) self-taught low-rank representation, (3) multi-view anomaly detection, (4) deeply refined robust feature.

Motion Detection. The aim of motion detection is to segment moving objects from the rest of the image [25]. Generally, popular motion detection methods includes *Background Subtraction* [26], [27], [28], *Temporal Differencing* [29], [30], and *Optical Flow* [31], [32], [33].

Object Classification. To transform motion detection to high-level representation, e.g., human, object and identify each of them for tracking in the later step, we need to classify current detected areas. There are mainly three categories: shape based classification [34], [35], motion-based classification [36], [37], and other methods [38], [39].

Object Tracking. Tracking [40], [41] is critical in surveillance system, e.g., tracking across distributed camera systems [42], in highly congested areas with crowds of people [43], tracking using mobile platforms [44], multi-sensor [45], [46], algorithm-fusion [47], feature integration [48].

In our framework, we will use both local and global features for anomaly event detection. In vision system, tracking and detection are "chicken and egg" problems, meaning solving one problem will help another. We therefore resort to a "tracking by detection" framework [49] that handles detection and tracking problems simultaneously in a unified framework. Then 3D HOG [50] and HOF [51] to are exploited to extract the low-level feature, which encodes the action or behavior of single or local units. For the global/crowd movements, we ignore the integration requirement of the trajectories and only focus on all the trajectories detected in the video and exploit anomaly detection of crowd behavior, e.g., moving crowd.

C. Self-Taught Low-Rank Representation

In this paper, we will tackle the multi-view anomaly detection problem from the perspective of data representation [52], [53]. In particular, for the sample set $X^{(i)}$ observed in the i-th view, we can represent it as:

$$X^{(i)} = X^{(i)}Z^{(i)} + E^{(i)}, \tag{1}$$

where $Z^{(i)}$ is a coefficient matrix and $E^{(i)}$ is a noise matrix. Assuming $X^{(i)}$ is drawn from c different classes, then the coefficient matrix $Z^{(i)}$ is expected to be low-rank. In other words, the coefficient vectors corresponding to samples within the same class should be highly correlated. For outliers that can not be reasonably presented by data drawn from c classes, they will be ruled out as "sample-specific" noise, which can be captured by the $l_{2,1}$ norm of matrix $E^{(i)}$. Since the anomaly representation $Z^{(i)}$ and $E^{(i)}$ is learned from data itself, we call it "self-taught low-rank representation".

Introducing the rank and sparse objectives on $Z^{(i)}$ and $E^{(i)}$, respectively, problem in Eq. (1) can be solved by Augment Lagrangian Multiplier (ALM) method in an iterative way [54], [52], meaning each time we update one unknown variable but fix the rest. In our large-scale anomaly detection problems, however, the computational complexity of ALM is prohibitively high. In our system, we will implement fast solution of the proposed low-rank and sparse modeling problem by recently published accelerated numerical approaches [55], [56].

D. Multi-View Anomaly Detection

The low-rank and sparse modeling enable to find the new representation Z and noise part E from self-taught fashion. This leads to a new formulation to detect different anomaly events required by different systems. Take two-view data for example, we will formulate the detection as:

$$\text{score}_k = Z_k^{(i)T} Z_k^{(j)} - \lambda E_k^{(i)T} E_k^{(j)}, \tag{2}$$

where i, j indicates two different views, score_k is an anomaly indicator, Z_k and E_k are low-rank coefficient and sparse noise for the k-th sample, and λ is a balancing parameter.

The above design benefits us to detect anomaly events of different levels. First, if the data is slightly perturbed or contaminated, then it will only change the local geometry of the neighborhood, but the overall underlying distribution of data will not change too much. Such anomaly can be successfully captured by the first term composed of Z_k. Second, if data are heavily contaminated, both $E_k^{(i)}$ and $E_k^{(j)}$ have large magnitude and their inner product is large as well. Therefore, the overall score will be significantly decreased. By setting different threshold, the proposed system can detect anomaly events of different levels:

- Normal event: large $Z_k^{(i)T} Z_k^{(j)}$, small $E_k^{(i)T} E_k^{(j)}$
- Subtle anomaly event: small $Z_k^{(i)T} Z_k^{(j)}$, small $E_k^{(i)T} E_k^{(j)}$
- Significant anomaly event: small $Z_k^{(i)T} Z_k^{(j)}$, large $E_k^{(i)T} E_k^{(j)}$

E. Deep Feature Refinement

Good feature representation, especially semantics, is essential to anomaly detection in video surveillance. In this section, we introduce how we refine anomaly representation by a deep structure. The basic principle of deep learning is to use multiple levels of representation of increasing complexity where the feature is abstracted and confined from the lower layer to the higher layer [57], [58], [59]. Inspired by this thought, we also build a deep structure for high-level representation learning and use the auto-encoder [58] as our building block, whose hidden layer can highly abstract the appearance from the lower layer.

To facilitate our large-scale video analytics, we propose to use marginalized denoising auto-encoder (MDA) to speedup deep feature learning [60]. Denoising auto-encoder is robust version of conventional auto-encoder by intentionally dropping out certain features in the output, and therefore trained a robust model. MDA further simplifies the auto-encoder by a single layer structure, which achieves a better balance between speed and performance. Similar to other deep models, MDA returns a weight matrix W that minimizes difference between original feature and contaminated feature. Then, the refined feature in the i-th view can be reformulated as: $W^{(i)}X^{(i)}$. Combining this with problem in Eq. (1), we obtain the proposed deeply refined robust feature in the i-th view as:

$$W^{(i)}X^{(i)} = W^{(i)}X^{(i)}Z^{(i)} + E^{(i)}. \tag{3}$$

By jointly solving W, Z and E, we could progressively improve the quality of anomaly representation.

V. Evaluation Framework

There are several existing databases for evaluating UAV video data exploitation that enables quick response or alert of an event. To name a few:

VIRAT[2]**.** The Video and Image Retrieval and Analysis Tool (VIRAT) program is a video surveillance project funded by the Information Processing Technology Office (IPTO) of the Defense Advanced Research Projects Agency (DARPA). It includes 8.5 hours HD videos of 12 different events and 11 outdoor scenes. VIRAT focuses heavily on developing means to be able to search through databases containing thousands of hours of video, looking for footage where certain types of activities took place, such as:

Person-to-Person: Following, meeting, gathering, moving as a group, dispersing, shaking hands, kissing, exchanging objects, kicking, carrying an object together.

Person-to-Vehicle: Driving, getting-in (out), loading (unloading), opening (closing) trunk, crawling under car, breaking window, shooting/launching, exploding/burning, dropping off, picking up.

UAV Dataset of PGB. UAV Dataset of Pedestrian Group Behavior simulates possible situations in pedestrian crowds which is captured by UAV [61]. Volunteers in the collection simulate predefined scenarios with minimal information provided. In this way, they could behave in a more natural way.

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

(A) VIRAT Dataset

(B) UAV Dataset of PGB

Fig. 4. Sample illustrations of (A) VIRAT and (B) PGB dataset.

The total duration of flight is 12 minutes, including videos of 15 different behaviors, e.g., parallel group motion, diverging, converging, etc. Samples of both datasets above can be found in Figure 4.

Data generation. As there are ground truth labels in these databases, we could directly use them for evaluations. In addition, to generate multi-view data, we randomly select a portion of feature vectors as one view, and repeat this n times for n views. Moreover, the data corruption can be simulated by randomly dropping a few features, or adding Gaussian noises.

REFERENCES

[1] M. Vidulich, C. Dominguez, E. Vogel, and G. McMillan, "Situation awareness: Papers and annotated bibliography. armstrong laboratory, crew systems directorate. wright-patterson afb oh," AL/CF-TR-1994-0085, Tech. Rep., 1994.

[2] F. T. Durso, T. R. Truitt, C. A. Hackworth, J. M. Crutchfield, D. Nikolic, P. M. Moertl, D. Ohrt, and C. A. Manning, "Expertise and chess: A pilot study comparing situation awareness methodologies," *Experimental analysis and measurement of situation awareness*, pp. 295–303, 1995.

[3] C. Drew, "Military is awash in data from drones," *New York Times*, vol. 10, p. 2010, 2010.

[4] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," in *IEEE Applied Imagery Pattern Recognition Workshop*. IEEE, 2008, pp. 1–8.

[5] H. Dee and D. Hogg, "Detecting inexplicable behaviour," in *British Machine Vision Conference*, vol. 477, 2004, p. 486.

[6] S. Kwak and H. Byun, "Detection of dominant flow and abnormal events in surveillance video," *Optical Engineering*, vol. 50, no. 2, pp. 027 202–027 202, 2011.

[7] X. Zhu, Z. Liu, and J. Zhang, "Human activity clustering for online anomaly detection," *Journal of Computers*, vol. 6, no. 6, pp. 1071–1079, 2011.

[8] A. Hendel, D. Weinshall, and S. Peleg, "Identifying surprising events in videos using bayesian topic models," *Asian Conference on Computer Vision*, pp. 448–459, 2011.

[9] I. Tziakos, A. Cavallaro, and L. Xu, "Local abnormality detection in video using subspace learning," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2010, pp. 519–525.

[10] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2871–2878.

[11] X. Wang, X. Ma, and W. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[12] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[13] J. Candamo, M. Shreve, D. Goldgof, D. Sapper, and R. Kasturi, "Understanding transit scenes: a survey on human behavior-recognition algorithms," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 1, pp. 206–224, 2010.

[14] J. Black, T. Ellis, P. Rosin *et al.*, "A novel method for video tracking performance evaluation," in *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003, pp. 125–132.

[15] L. Brown, A. Senior, Y. Tian, J. Connell, A. Hampapur, C. Shu, H. Merkl, and M. Lu, "Performance evaluation of surveillance systems under varying conditions," in *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. Citeseer, 2005, pp. 1–8.

[16] S. Muller-Schneiders, T. Jager, H. Loos, and W. Niem, "Performance evaluation of a real time video surveillance system," in *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 137–143.

[17] D. Doermann and D. Mihalcik, "Tools and techniques for video performance evaluation," in *International Conference on Pattern Recognition*, vol. 4. IEEE, 2000, pp. 167–170.

[18] F. Yin, D. Makris, and S. Velastin, "Performance evaluation of object tracking algorithms," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2007.

[19] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319–336, 2009.

[20] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *European Conference on Computer Vision*. Springer, 2012, pp. 808–821.

[21] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 113–126, 2014.

[22] M. Fiori, P. Sprechmann, J. Vogelstein, P. Musé, and G. Sapiro, "Robust multimodal graph matching: Sparse coding meets graph matching," in *Advances in Neural Information Processing Systems*, 2013, pp. 127–135.

[23] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 1–42, 2014.

[25] C. Cedras and M. Shah, "A survey of motion analysis from moving light displays," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1994, pp. 214–221.

[26] J. Heikkila and O. Silvén, "A real-time system for monitoring of cyclists and pedestrians," in *IEEE Workshop on Visual Surveillance*. IEEE, 1999, pp. 74–81.

[27] R. Cutler and L. Davis, "View-based detection and analysis of periodic motion," in *International Conference on Pattern Recognition*, vol. 1. IEEE, 1998, pp. 495–500.

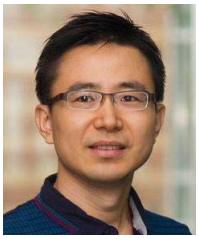
[28] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, 2005.

[29] F. Hu, Y. Zhang, and L. Yao, "An effective detection algorithm for moving object with complex background," in *International Conference on Machine Learning and Cybernetics*, vol. 8. IEEE, 2005, pp. 5011–5015.

[30] Y. Choi, P. Zaijun, S. Kim, T. Kim, and C. Park, "Salient motion information detection technique using weighted subtraction image and motion vector," in *International Conference on Hybrid Information Technology*, vol. 1. IEEE, 2006, pp. 263–269.

[31] B. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," in *international joint conference on Artificial intelligence*, 1981.
[32] M. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer vision and image understanding*, vol. 63, no. 1, pp. 75–104, 1996.
[33] R. Szeliski and J. Coughlan, "Spline-based image registration," *International Journal of Computer Vision*, vol. 22, no. 3, pp. 199–218, 1997.
[34] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
[35] B. Chee, M. Lazarescu, and T. Tan, "Detection and monitoring of passengers on a bus by video surveillance," in *International Conference on Image Analysis and Processing*. IEEE, 2007, pp. 143–148.
[36] D. Gavrila, "The visual analysis of human movement: A survey," *Computer vision and image understanding*, vol. 73, no. 1, pp. 82–98, 1999.
[37] C. Cedras and M. Shah, "Motion-based recognition a survey," *Image and Vision Computing*, vol. 13, no. 2, pp. 129–155, 1995.
[38] S. Kim and H. Kim, "Face detection using multi-modal information," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2000, pp. 14–19.
[39] S. Harasse, L. Bonnaud, and M. Desvignes, "Human model for people detection in dynamic scenes," in *International Conference on Pattern Recognition*, vol. 1. IEEE, 2006, pp. 335–354.
[40] S. Haykin and N. deFreitas, "Special issue on sequential state estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 399–400, 2004.
[41] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm Computing Surveys (CSUR)*, vol. 38, no. 4, p. 13, 2006.
[42] N. Ning and T. Tan, "A framework for tracking moving target in a heterogeneous camera suite," in *International Conference on Control, Automation, Robotics and Vision*. IEEE, 2006, pp. 1–5.
[43] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.
[44] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.
[45] U. Scheunert, H. Cramer, B. Fardi, and G. Wanielik, "Multi sensor based tracking of pedestrians: A survey of suitable movement models," in *IEEE Intelligent Vehicles Symposium*. IEEE, 2004, pp. 774–778.
[46] G. Foresti, C. Regazzoni, and P. Varshney, *Multisensor surveillance systems: the fusion perspective*. Springer, 2003.
[47] N. Siebel and S. Maybank, "Fusion of multiple tracking algorithms for robust people tracking," *European Conference on Computer Vision*, pp. 373–387, 2006.
[48] B. Leibe, K. Schindler, and L. Van Gool, "Coupled detection and trajectory estimation for multi-object tracking," in *International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.
[49] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *International Conference on Computer Vision*. IEEE, 2009, pp. 1515–1522.
[50] A. Klaser and M. Marszalek, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, 2008.
[51] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.
[52] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *International Conference on Machine Learning*, 2010, pp. 663–670.
[53] S. Li, M. Shao, and Y. Fu, "Multi-view low-rank analysis for outlier detection," in *SIAM International Conference on Data Mining (SDM)*, 2015.
[54] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.
[55] S. Xiao, W. Li, D. Xu, and D. Tao, "Falrr: A fast low rank representation solver," June 2015.
[56] C. Li, L. Lin, W. Zuo, S. Yan, and J. Tang, "Sold: Sub-optimal low-rank decomposition for efficient video segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
[57] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
[58] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.
[59] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.
[60] M. Chen, Z. Xu, F. Sha, and K. Q. Weinberger, "Marginalized denoising autoencoders for domain adaptation," in *Proceedings of the 29th International Conference on Machine Learning*, 2012, pp. 767–774.
[61] F. Burkert and F. Fraundorfer, "Uav-based monitoring of pedestrian groups," *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, no. 2, pp. 67–72, 2013.

Ming Shao received the B.E. degree in computer science, the B.S. degree in applied mathematics, and the M.E. degree in computer science and engineering from Beihang University, Beijing, China, in 2006, 2007, and 2009, respectively. He is currently working toward the Ph.D. degree in Department of Electrical and Computer Engineering at Northeastern University, Boston, MA. His current research interests include sparse modeling, low-rank matrix analysis, and applied machine learning on social media analytics. He was the recipient of the Presidential Fellowship of State University of New York at Buffalo from 2010 to 2012, and the best paper award winner of IEEE ICDM 2011 Workshop on Large Scale Visual Analytics. He has served as the reviewers for IEEE journals: IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, etc. He is an IEEE student member.



Yun Fu (S'07-M'08-SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. He is an interdisciplinary Associate Professor affiliated with College of Engineering and the College of Computer and Information Science at Northeastern University since 2012. Dr. Fu's research interests are Interdisciplinary research in Machine Learning, Social Media Analytics, Human-Computer Interaction, and Cyber-Physical Systems. His research is broadly supported by NSF, DOD, DARPA, IARPA, ONR, AFOSR, ARL/ARO, NGA, IC, and Google. He serves as Associate Editor of the IEEE Transactions on Neural Networks and Leaning Systems (TNNLS), and IEEE Transactions on Circuits and Systems for Video Technology (TCSVT).