AUTOMATIC SCENE UNDERSTANDING SBIR PHASE II, TOPIC NO. OSD10-L04 REPORT NO. 1, NOVEMBER 30, 2012

[1]*Lam Tran*, [2]*Carmeliza Navasca*, [1]*Jiebo Luo*

[1]University of Rochester, [2]University of Alabama at Birmingham
lam.tran@rochester.edu, cnavasca@uab.edu, jluo@cs.rochester.edu

ABSTRACT

In this report, we investigate a method for anomaly detection in surveillance video in a tensor framework. We treat a video as a tensor and utilize a stable PCA to decompose it into two tensors, the first tensor is a low rank tensor that consists of background pixels and the second tensor is a sparse tensor that consists of the foreground pixels. The sparse tensor is then analyzed to detect anomaly. The proposed method is a one-shot framework to determine frames that are anomalous in a video.

***Index Terms*—** Anomaly detection, Surveillance Video, Stable PCA, and Tensor.

1. INTRODUCTION

There are a growing interest in anomaly or unusual detection in computer vision and other fields. In computer vision, Zhao et al. [1] proposed to detect anomalous events in video based on a sparse code reconstruction of an dynamic dictionary. Their method uses the first segment of a video to build a dictionary and dynamically updating the dictionary for new observations. An unusual event is flagged when the reconstruction error is larger than a pre-defined threshold.

Benezeth et al. [2] incorporated spatio-temporal co-occurrences for abnormal event detection. A simple background substraction is first applied to a video to extract the motion and spatial position of foreground objects. An MRF model distribution accounts for speed, size, and position of the object are combined in learning co-occurrence relationship. This distribution is then used to classify anomalous event based on a pre-defined threshold.

Our work follows the direction of Wen et al. [3] and mainly motivated by Sun et al. [4]. Sun introduced several tensor decomposition methods for anomaly detection in network monitoring based on network flows. Wen used the same framework proposed by Sun for anomaly detection in surveillance video. The work proposed by Wen is based on a background modeling that is obtained by low dimensional tensor decomposition. The anomaly detection is alarmed when the reconstruction error passed an threshold.

It is difficult for us to do a survey on anomaly detection and tensor decompositions in this short report. We introduced papers that are most relevant with our work. We refer the reader to the work of [5, 6, 7] for tensor decompositions and applications. The remainder of report is organized as follows: section 2 introduces preliminaries of mathematics, section 3 presents our model, section 4 shows numerical results, and section 5 concludes this report.

2. PRELIMINARIES

We denote the scalars in $\mathbb{R}$ with lower-case letters $(a, b, \ldots)$ and the vectors with bold lower-case letters $(\mathbf{a}, \mathbf{b}, \ldots)$. The matrices are written as bold upper-case letters $(\mathbf{A}, \mathbf{B}, \ldots)$ and the symbol for tensors are calligraphic letters $(\mathcal{A}, \mathcal{B}, \ldots)$. The subscripts represent the following scalars: $(\mathcal{A})_{ijk} = a_{ijk}$, $(\mathbf{A})_{ij} = a_{ij}$, $(\mathbf{a})_i = a_i$. The superscripts indicate the dimension size.

A matrix $\mathbf{T} \in \mathbb{R}^{I \times J}$ is a second order tensor which has an SVD of $\mathbf{T} = \mathbf{U\Sigma V}'$ where $\mathbf{\Sigma} = diag\{\sigma_1, \sigma_2, \cdots, \sigma_p\}$ where $p = \min\{I, J\}$ and $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices. The Frobenius norm of $\mathbf{T}$ is defined as $\|\mathbf{T}\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2}$ while the trace class norm is the sum of its singular values, i.e. $\|\mathbf{T}\|_* = \sigma_1 + \sigma_2 + \cdots + \sigma_p$. Moreover, the 1-norm of $\mathbf{T}$ is $\|\mathbf{T}\|_1 = \sum_{ij} |\mathbf{T}_{ij}|$. A rank r matrix $\widehat{\mathbf{T}}$ of $\mathbf{T}$, i.e. $r < p = rank(\mathbf{T})$, is defined as $\widehat{\mathbf{T}} = \mathbf{U}_r \mathbf{S}_r \mathbf{V}'_r = \sum_{i=1}^r \sigma_i \cdot u_i v_i'$.

2.1. Tensor Basics and Tensor SVD

The order of a tensor refers to the cardinality of the index set. A matrix is a second-order tensor and a vector is a first-order tensor.

Definition 2.1 (Tucker mode-n product) *Given a tensor $\mathcal{T} \in \mathbb{R}^{I \times J \times K}$ and the matrices $\mathbf{U_1} \in \mathbb{R}^{\hat{I} \times I}$, $\mathbf{U_2} \in \mathbb{R}^{\hat{J} \times J}$ and $\mathbf{U_3} \in \mathbb{R}^{\hat{K} \times K}$, then the Tucker mode-$n$ products are as follows: $(\mathcal{T} \bullet_1 \mathbf{U_1})_{\hat{i},j,k} = \sum_{i=1}^I \mathcal{T}_{ijk} \mathbf{A}_{\hat{i}i}$ (mode-1 product), $(\mathcal{T} \bullet_2 \mathbf{U_2})_{\hat{j},i,k} = \sum_{j=1}^J \mathcal{T}_{ijk} \mathbf{B}_{\hat{j}j}$ (mode-2 product) and $(\mathcal{T} \bullet_3 \mathbf{U_3})_{\hat{k},i,j} = \sum_{k=1}^K \mathcal{T}_{ijk} \mathbf{C}_{\hat{k}k}$, (mode-3 product).*

The tensor SVD is also referred to multilinear SVD (or higher-order SVD).

Theorem 2.1 (Multilinear SVD [8]) *A third order tensor $\mathcal{T} \in \mathbb{R}^{I\times J\times K}$ can be represented as a product*

$$\mathcal{T} = \mathcal{S} \bullet_1 \mathbf{U_1} \bullet_2 \mathbf{U_2} \bullet_3 \mathbf{U_3}$$

where $\mathbf{U_1} \in \mathbb{R}^{I\times I}$, $\mathbf{U_2} \in \mathbb{R}^{J\times J}$ and $\mathbf{U_3} \in \mathbb{R}^{K\times K}$ are orthogonal matrices. The core tensor $\mathcal{S} \in \mathbb{R}^{I\times J\times K}$ are the matricized subtensors $\mathbf{S}^1_{i=\alpha} \in \mathbb{R}^{J\times K}$, $\mathbf{S}^2_{j=\alpha} \in \mathbb{R}^{I\times K}$ and $\mathbf{S}^3_{k=\alpha} \in \mathbb{R}^{I\times J}$ with the following properties:

- *all-orthogonality:*
  $\langle \mathbf{S}^n_{i_n=\alpha}, \mathbf{S}^n_{i_n=\beta}\rangle = (\sigma^{(1)}_\alpha)^2 \delta_{\alpha,\beta}, \ \ \alpha,\beta = 1,\ldots,I_n,$
- *ordering:*
  $\|\mathbf{S}^n_{i_n=1}\|_F \geq \|\mathbf{S}^n_{i_n=2}\|_F \geq \cdots \geq \|\mathbf{S}^n_{i_n=I_n}\|_F \geq 0$

where $\|\mathbf{S}^n_{i_n=\alpha}\|_F = \sigma^{(n)}_\alpha$ for $\alpha = 1,\ldots,I_n$ ($I_1 = I, I_2 = J, I_3 = K$ and $i_1 = i, i_2 = j, i_3 = k$).

The usual inner product of matrices, $\mathbf{A},\mathbf{B} \in \mathbb{R}^{I\times J}$ is denoted by $\langle \mathbf{A},\mathbf{B}\rangle = \sum_{ij} b_{ij}a_{ij}$. For a third order tensor, there are three sets of singular values: $\sigma^{(1)}_\alpha$'s are mode-1 singular values, $\sigma^{(2)}_\alpha$'s are the mode-2 singular values and $\sigma^{(3)}_\alpha$'s are the mode-3 singular values. The corresponding mode-1, mode-2 and mode-3 singular vectors are $\mathbf{u_1}_\alpha$, $\mathbf{u_2}_\alpha$ and $\mathbf{u_3}_\alpha$, respectively.

The following matrix representations of tensor SVD is obtained by unfolding the third-order $\mathcal{T}$ and $\mathcal{S}$ tensors in [8]:

$$\begin{aligned} \mathbf{T}^1 &= \mathbf{U_1}\mathbf{S}^1(\mathbf{U_2}\otimes\mathbf{U_3})^T, \ \ \mathbf{T}^2 = \mathbf{U_2}\mathbf{S}^2(\mathbf{U_3}\otimes\mathbf{U_1})^T, \\ \mathbf{T}^3 &= \mathbf{U_3}\mathbf{S}^3(\mathbf{U_1}\otimes\mathbf{U_2})^T \end{aligned} \tag{1}$$

We denote $\mathbf{T}^1 = \mathbf{T}^{I\times JK}$, $\mathbf{T}^2 = \mathbf{T}^{J\times KI}$, $\mathbf{T}^3 = \mathbf{T}^{K\times IJ}$ and similarly for $\mathbf{S}^n$. In general, $\mathbf{T}^n$ and $\mathbf{S}^n$ are mode-n matrix representation of $\mathcal{T}$ and $\mathcal{S}$.

3. MODEL

In surveillance video, stationary cameras are often used to monitor the scene for security threats. It is normal for us to assume that the background pixels remain unchanged throughout the video. This enable us to use only the foreground pixels to detect frames that are anomalous unlike the work proposed by [3] which relies on the background pixels.

In our model, we define a frame as anomalous when the activities in the current frame standout when compared to the activities in the previous frames. Given that we have N frames in the video, we only consider the most current R frames to determine whether the current frame is anomalous. The number of frames used to determine whether the current frame is anomalous is a small fraction of the total number frames in the video. This approach analogous to having a forgetting factor that weights more on recent activities and less on past activities.

3.1. Video Representation

The proposed method represents a video as a tensor $\mathcal{X}$ and find the low multi-linear rank $\mathcal{X}^{(i)}_L$ and sparse $\mathcal{X}^{(i)}_S$ tensors for mode $i = 1,2,3$. The process of converting a tensor $\mathcal{X} \in \mathbb{R}^{I\times J\times K}$ to matrix $\mathbf{X}^1 \in \mathbb{R}^{JK\times I}$ is known as the *unfolding* process. The reverse process of converting a matrix back into a tensor is known as the folding process. There are more than one way to arrange the columns of the matrix when *unfolding* a tensor. Our method *unfolds* a tensor into a matrix by slicing a tensor into slices where each slice represent a matrix. Each slice is then vectorized into a column vector and stores in the matrix in chronological order.

For example, a video can be represented as a tensor $\mathcal{X} \in \mathbb{R}^{I\times J\times K}$, where I and J are the spatial dimensions of a video frame and K is the total number of frames. *Unfolding* the tensor $\mathcal{X}$ into mode-3 matrix $\mathbf{X}^3 \in \mathbb{R}^{IJ\times K}$ is done by slicing out one frame at a time, vectorizes each frame into a column vector and store each vector at the column corresponding to its frame position of the tensor. This setup easily obtain an analysis of how spatial information change in time in our model. These *unfolded* tensors satisfy the equations in (1) w.r.t. SVD.

In comparison to the work of [1], which computes salient feature points of a video by using the cuboid feature points extractor [9], we use the foreground pixels as our salient feature data. The feature data in our method is more reliable than cuboid points. This is due to each cuboid point is a 3D moving corner and they are rare in a video. Moreover, cuboid points are inadequate to represent anomalous activities in a video since each cuboid point only captures part of an activity. Our main idea is to extract the foreground pixels accurately and analyze the foreground pixels to determine the frames that are anomalous.

3.2. Stable Principal Component Pursuit

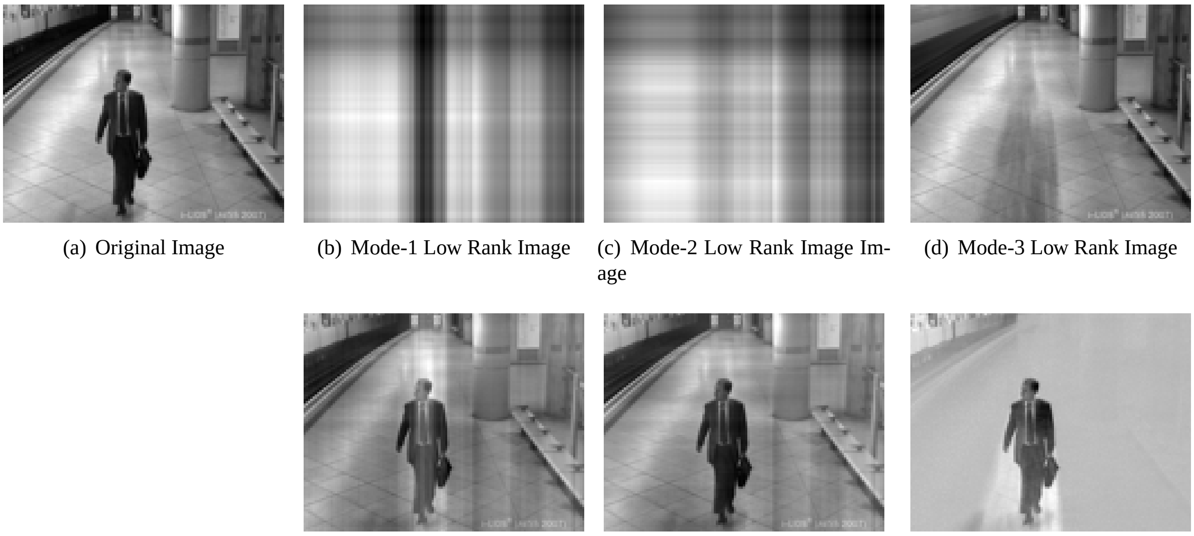
Stable principal component pursuit PCP [10] is a convex programming method for separating a tensor data into a sum of low rank and sparse frames. The following optimization,

$$\begin{aligned} &\text{minimize } \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \\ &\text{subject to } \mathbf{M} = \mathbf{L} - \mathbf{S} \end{aligned}$$

achieves the separation. The algorithm above is based on an augmented Lagrange multiplier ALM [11, 12] with the augmented Lagrangian as

$$l(\mathbf{L},\mathbf{S},\mathbf{Y}) = \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \langle \mathbf{Y}, \mathbf{M}-\mathbf{L}-\mathbf{S}\rangle + \frac{\mu}{2}\|\mathbf{M}-\mathbf{L}-\mathbf{S}\|^2_F$$

with multiplier $\mathbf{Y}$. The steps in the algorithm coincide with the alternating minimization of $l(\mathbf{L},\mathbf{S},\mathbf{Y})$ w.r.t. $\mathbf{L}$ and $\mathbf{S}$ while also updating $\mathbf{Y}$. In Fig. 1, let $\mathcal{S}_{\frac{\lambda}{\mu}} : R \rightarrow R$ is the shrinkage operator defined as $\mathcal{S}_{\frac{\lambda}{\mu}} = \text{sgn}(x)\max(|x| - \frac{\lambda}{\mu}, 0)$ and $\mathcal{D}_{\lambda^{-1}}(\mathbf{X}) = \mathbf{U}\mathcal{S}_{\lambda^{-1}}(\mathbf{\Sigma})\mathbf{V}'$ with $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ is the singular value thresholding operator.



(a) Original Image (b) Mode-1 Low Rank Image (c) Mode-2 Low Rank Image Image (d) Mode-3 Low Rank Image

(e) Mode-1 Sparse Image (f) Mode-2 Sparse Image (g) Mode-3 Sparse Image

Fig. 2. An illustration of PCP decomposition of each mode. Note that each of the low rank and sparse matrices shown above are reshaped for demonstration purpose. Beside mode-3, the other two modes have no visual interpretation.

```
Algorithm 1 (Principal Component Pursuit by Alternating Directions
initialize: $S_0 = Y_0 = 0, \mu > 0$.
while not converged do
    compute $L_{k+1} = \mathcal{D}_{\mu^{-1}}(M - S_k + \mu^{-1}Y_k)$;
    compute $S_{k+1} = \mathcal{S}_{\lambda\mu^{-1}}(M - L_{k+1} + \mu^{-1}Y_k)$;
    compute $Y_{k+1} = Y_k + \mu(M - L_{k+1} - S_{k+1})$;
end while
output: $L, S$.
```

Fig. 1. PCP

3.3. Low Multilinear Rank and Sparse Tensors

For each mode, we *unfold* the tensor into a mode-i matrix by taking one slice at a time in chronological order, satisfying (1). Each slice is then a matrix, and the matrtix is vectorized into a column vector to store it at a column in $\mathbf{X}^i$. Although the ordering is not important in general, we should point out that when dealing with time data, it is important to have the data in chronological order. Then for each matrix $\mathbf{X}^i$ we apply the PCP algorithm:

$$\begin{aligned} &\text{minimize } \|\mathbf{L}^i\|_* + \lambda\|\mathbf{S}^i\|_1 \\ &\text{subject to } \mathbf{X}^i = \mathbf{L}^i - \mathbf{S}^i \end{aligned}$$

The low rank tensors are discarded after the tensors decomposition step.

To detect anomaly, we calculate the mean of Frobenius norms of each frame t of $\mathcal{X}_S^i$ and apply the following criteria to detect anomaly:

$$\sum_{i=1}^{3} e_t^i \leq \sum_{i=1}^{3} \left[\text{mean}\left(e_j^i|_{j=t-R}^t\right) + 2 \cdot \text{std}\left(e_j^i|_{j=t-R}^t\right)\right], \quad (2)$$

where

$$e_t^i = \|\mathcal{X}_S^{i,t}\|_F^2.$$

When the error of a frame is greater than two standard deviations from the mean, the frame is an outlier and labelled as an anomalous frame. We only used the previous R frames instead of all of the frames to determine whether the current frame is anomalous.

4. NUMERICAL EXPERIMENTS

We downloaded two videos from AVSS and ViSOR datasets to demonstrate our algorithm. We used 150 frames of a man walking down in a subway station and created a tensor by stacking up the frames. We then *unfolded* the tensor to 3 modes and applied PCP to decompose the data into a low rank and sparse matrices for each model. The matrices are *folded* back into sparse and lower rank tensors for each mode. An illustration of the PCP tensor decomposition is shown in Fig. 2. The top left image is an original frame in a video and the other columns are the decomposed into low rank and sparse matrices of the same frame of the video.

Our method is a one shot method for determining a set of frames that are anomalous based on the criteria of (2). Our result is illustrated in Fig. 4 of a man waiting for a subway to arrive. He then leaves the scene with his belonging behind. The proposed algorithm was able to determine this and labelled those frames as anomalous due to a sudden change in the foreground tensors.

The ViSOR video is captured in a laboratory of people walking around with a person opening a locker. A sample of these frames are illustrated in Fig. 5. The criteria of (2) is also used to determine outliers. Fig. 3 shows the Frobenius norm of each frame and the frames with Frobenius norm above the red line are consider outliers and anomalous. The frames in this sequence labelled as anomalous are shown in Fig. 6.

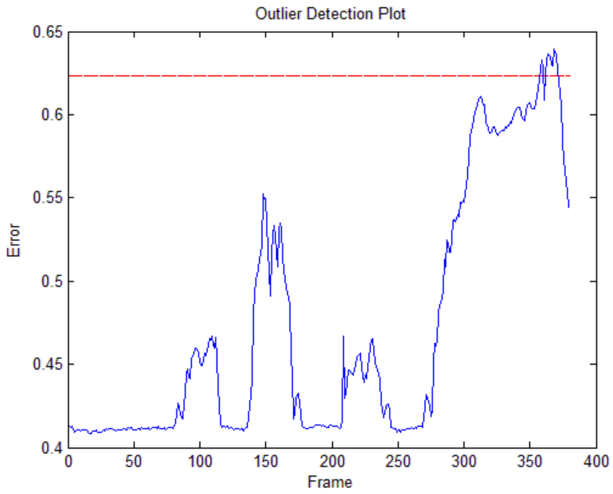


Fig. 3. Plot of Frobenius norm of sparse tensor of ViSOR data set. The frames above the red line are outlier and consider outlier and anomalous.

5. CONCLUSION

In this report, we proposed a video anomaly detection algorithm via low-rank and sparse decompositions. Our work focuses on extracting the foreground pixels accurately and analyze the foreground tensors for frames that are anomalous in a one shot framework. We plan to extend this framework to subspace anomaly detection for events instead of frames.

6. REFERENCES

[1] Bin Zhao, Li Fei-Fei, and Eric P. Xing, “Online detection of unusual events in videos via dynamic sparse coding,” in *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[2] Yannick Benezeth, Pierre-Marc Jodoin, Venkatesh Saligrama, and Christophe Rosenberger, “Abnormal events detection based on spatio-temporal cooccurences,” in *CVPR*. 2009, pp. 2458–2465, IEEE.

[3] Jing Wen Xinbo Gao Jie Li, Guan Han, “Robust tensor subspace learning for anomaly detection,” in *International Journal of Machine Learning and Cybernetics*, Colorado Springs, CO, 2011, pp. 89–98.

[4] Jimeng Sun, Dacheng Tao, and Christos Faloutsos, “Beyond streams and graphs: dynamic tensor analysis,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2006, KDD ’06, pp. 374–383, ACM.

[5] Tamara G. Kolda and Brett W. Bader, “Tensor decompositions and applications,” *SIAM REVIEW*, vol. 51, no. 3, pp. 455–500, 2009.

[6] C.Navasca and L. De Lathauwer, “Low Multilinear Rank Tensor Decomposition via Semidefinite Programming,” *17th Proceedings of the European Signal Processing Conference*, 2009.

[7] L. McGowan D. Burdick, X. Tu and D. Millican, “Resolution of Multicomponent Fluorescent Mixtures by Analysis of the Excitation-Emission-Frequency Array,” *Journal of Chemometrics*, 1990.

[8] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle, “A multilinear singular value decomposition,” *SIAM J. Matrix Anal. Appl*, vol. 21, pp. 1253–1278, 2000.

[9] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proceedings of the 14th International Conference on Computer Communications and Networks*, Washington,

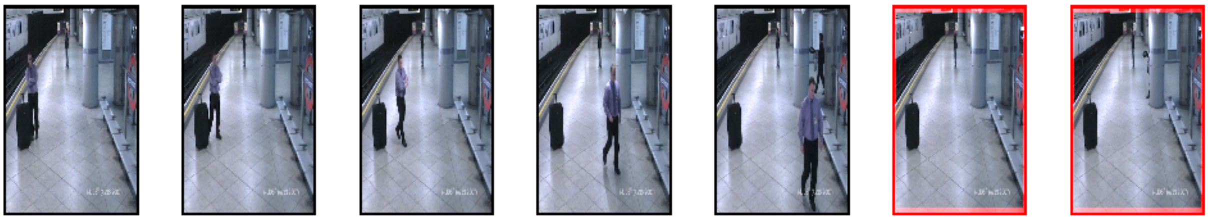


Fig. 4. The figures above are a sequence of frames of a man waiting for the subway to arrive. He leaves his belonging behind to exit the slight of the camera. Our anomaly detection algorithm observed this by analyzing the foreground tensors. The figures with red boundary are classified as anomalous.

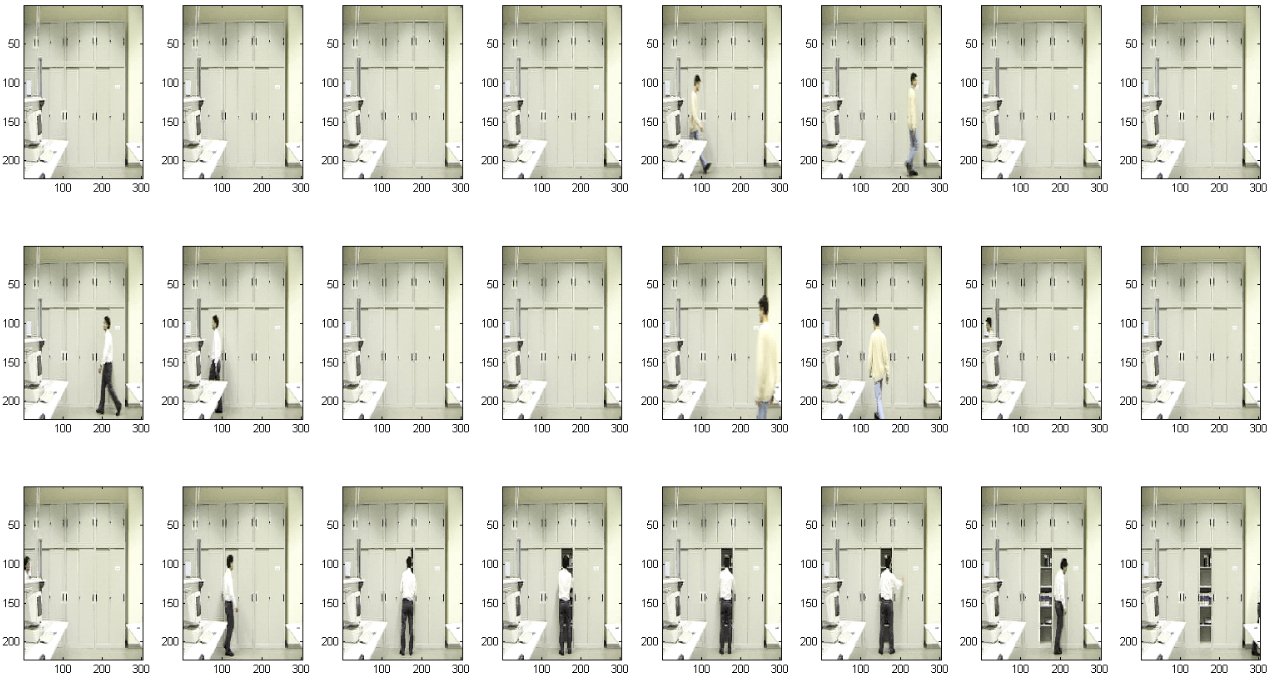


Fig. 5. The images above are a sample of frames from ViSOR dataset. It captured a lab of people walking around and a person open a lock to get items from the locker.

DC, USA, 2005, ICCCN ’05, pp. 65–72, IEEE Computer Society.

[10] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright, “Robust principal component analysis?,” *J. ACM*, vol. 58, no. 3, pp. 11, 2011.

[11] Zhouchen Lin, Minming Chen, and Yi Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” 2010.

[12] X. M. Yuan and J. Yang, “Sparse and low-rank matrix decomposition via alternating direction methods,” *Pacific Journal of Optimization*, 2009.

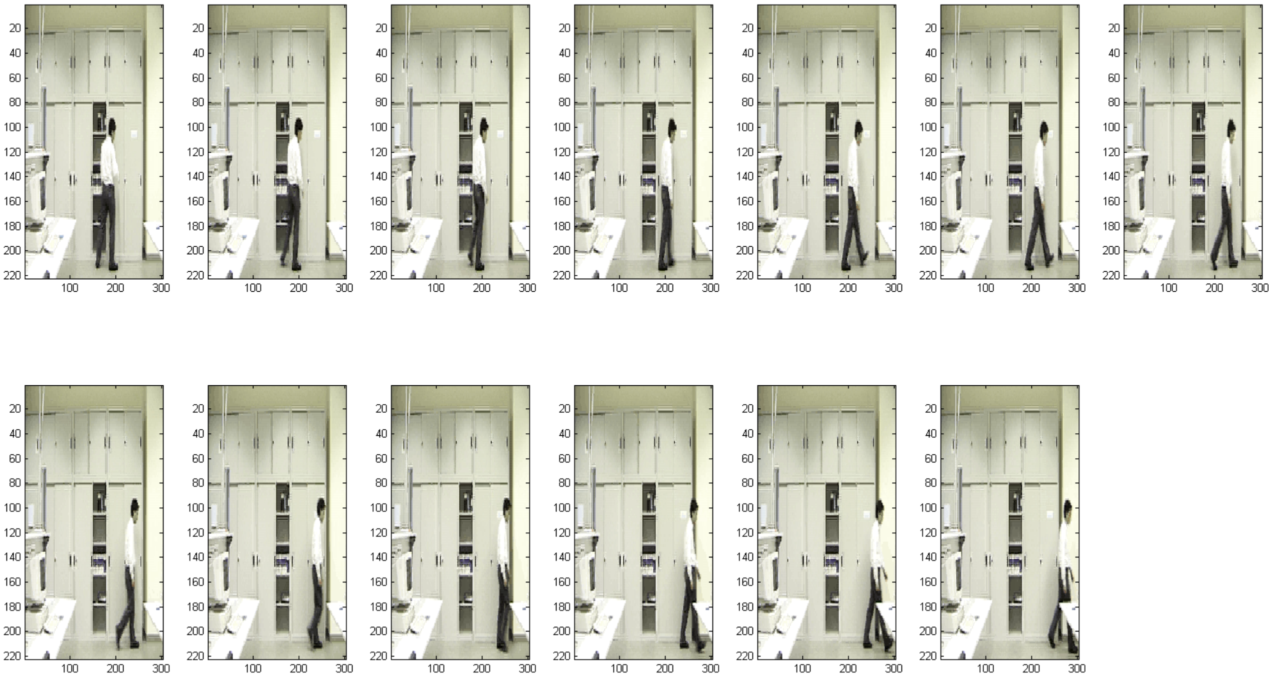


Fig. 6. The images above are a sample of frames from ViSOR dataset. It captured a lab of people walking around and a person open a lock to get items from the locker.