CHANGE DETECTION FOR AREA SURVEILLANCE USING A MOVING CAMERA

Tatsuhisa Watanabe, Tomoharu Nakashima, and Yoshifumi Kusunoki
Graduate School of Humanities and Sustainable System Sciences
Osaka Prefecture University
Gakuen-cho 1-1, Sakai, Osaka 599-8531, Japan
Email: {tatsuhisa.watanabe, tomoharu.nakashima, yoshifumi.kusunoki}@kis.osakafu-u.ac.jp

KEYWORDS

Change detection, Area surveillance, Monocular camera, Autonomous robot.

ABSTRACT

This paper tackles area surveillance with a moving camera by change detection. None of the existing datasets for change detection meets a surveillance scenario where a camera is mounted on a moving platform and pointed in the direction of moving. Thus, this paper creates a new dataset including several challenging points. For this dataset, this paper employs a composable method and proposes some components. To evaluate the proposed components, some corresponding classic methods were also tested on the dataset. As a result, the proposals outperformed them. Moreover, this paper investigated the relationship between the parameters of the components and their performance.

INTRODUCTION

Surveillance systems have been attracting attention because they have a great potential to reduce the workload of monitors. Surveillance systems can be applied to various fields such as urban monitoring, agriculture, and traffic analysis with manifold sensors. In recent years, some sensors have been mounted on Unmanned Aerial Vehicles or Unmanned Ground Vehicles (UGVs) as the development of industrial technologies. This paper aims to automatically observe areas for security using these autonomous vehicles. To do so, this paper overviews some previous methods and datasets in the following paragraphs and proposes a new dataset and method.

Regarding the area monitoring, there are two types of automatic methods: target-limited and target-agnostic. The target-limited methods focus on a specific anomaly including human behaviors (Morais et al. 2019; Singh et al. 2018). While the target-limited methods performed well, they cannot be a complete replacement for humans because they can only detect the expected target. The idea of combining them does not work because one cannot obtain or even list all possible abnormal patterns. The target-agnostic methods assume the available data as a distribution of normal situations and detect samples far from it as anomalies (Chu et al. 2019; Hao et al. 2019).

The target-agnostic methods for area surveillance can be roughly divided into two groups based on situation types. One group is for the place where people are NOT supposed to be. Methods of this sort have to be able to detect the emergence or disappearance of anything. The other group is for the place where people appear. In such places, detectors are required to report anomalous behaviors of humans too. The former situation can be solved by change detection (CD) and the latter by anomaly detection. The former is more important in practice because if there are people, they can take action.

Although a large number of studies devised CD methods, all of them employed fixed cameras. While research of this kind plays an important role in surveillance, blind spots of the fixed cameras can arouse a controversy over security. The blind spots can be reduced by mounting cameras on a moving platform such as autonomous vehicles.

There are only four datasets for CD with moving vehicles. One dataset, known as VDAO (Silva et al. 2014), was created inside an offshore facility. A camera on a mobile robotic platform on a straight rail was used. This dataset contains 15 different abnormal objects such as bags. (Sakurada and Okatani 2015) constructed two datasets consisting of panoramic images: TSUNAMI and GSV. TSUNAMI captured scenes of tsunami-damaged areas in Japan. GSV is a collection of images on Google Street View. The last dataset is the so-called VL-CMU-CD dataset (Alcantarilla et al. 2018). It includes pictures taken in the city of Pittsburgh, PA, USA, over a year.

The four CD datasets with moving devices do not suffice for area surveillance. Every dataset but VDAO (i.e. TSUNAMI, GSV, and VL-CMU-CD) was not designed for the field of surveillance. Thus, their change types such as buildings are not anomalous. The VDAO dataset only contains abandoned objects, not humans. Moreover, they do not contain looming motion. Such motion is unavoidable if one employs a moving camera in a narrow place including a hallway. In response to the inadequacy of the existing datasets, a new CD dataset is created with a moving monocular camera on a hallway. Compared to the existing datasets, the proposed one has some challenging points: 1) looming motion, under which everything varies grad-

ually in size and position; 2) non-identical trajectory and inconsistent viewing angles, which cause parallax leading to false alarm; 3) illumination change due mainly to different times of the day. The dataset is detailed more in the EXPERIMENTS AND RESULTS section. To tackle the proposed dataset, a composable procedure is employed as in (Carvalho et al. 2019). However, it was proposed for the VDAO dataset, so three new components are designed for the proposed dataset: Video Compression (VC), Temporal Alignment (TA), and Frame Comparison (FC).

To evaluate the proposed TA and FC components, classic TA and FC methods are also tested. (Evangelidis and Bauckhage 2013) proposed a TA method using local descriptors. One of the TA datasets they tackled is similar to the proposed dataset, which contains looming motion. (Carvalho et al. 2019) proposed a structured method for the VDAO dataset and employed Zero-mean Normalized Cross Correlation (ZNCC) for dissimilarity calculation.

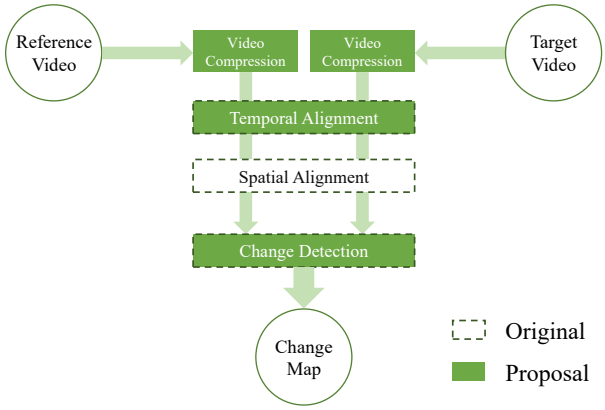
The contributions of this paper are two-fold.

- A dataset has been created with a moving monocular camera. This is the first video-CD dataset that has looming-motion for area surveillance.
- A structured way has been proposed to deal with the proposed dataset.

METHOD

This paper employs a composable procedure as in (Carvalho et al. 2019) with three new components. Figure 1 illustrates its whole procedure, where dotted line boxes indicate that the corresponding part was originally proposed in the literature, and filled-in boxes are the new components that this paper proposes. The input is a pair of videos: reference video and target one. The method detects changes in the target video against the reference one. The first process, VC, is a novel process for reducing the computation cost of the downstream processes. TA synchronizes a compressed version of the target video to that of the reference one. SA performs image registration between each target frame and the matched reference one. FC computes dissimilarity values of each matched frame pair and binarizes them with a threshold.

For each component, visual information of videos needs to be extracted. To do so, famous CNN architectures, called VGG13 and VGG16 (Simonyan and Zisserman 2014), are used. They consist of three types of layers: convolution, max pooling, and fully connected layer. In this paper, all the fully connected layers of VGGs are discarded because the components demand just spatial information, not features for classification. Training a deep network requires a considerable amount of data and time. Therefore, this paper exploits the pre-trained parameters of VGGs on the ImageNet dataset (Deng et al. 2009). For VC and TA, the last pooling layer of VGG16 is replaced with global average pooling (GAP) (Lin et al. 2013). GAP is considered to lose spatial information. However, it seems that this weakness potentially renders a matching method insensitive to parallax or too partial textures. The GAP version of VGG16, noted as VGG16', returns a C dimensional feature map for the input image.



Figures 1: The Whole Procedure of the Proposed Method

Proposed Video Compression (VC)

With the aim of area surveillance, it is not necessary to inspect all frames since several consecutive frames contain almost the same contents. Therefore, VC reduces redundant frames according to the similarity of sequential ones. Let $S^r = \{I^r_1, I^r_2, ..., I^r_{n_r}\}, S^t = \{I^t_1, I^t_2, ..., I^t_{n_t}\}$ be the reference sequence and the target one, respectively. n_r and n_t are the number of the reference frames and that of the target ones, respectively. The aim of this component is to retrieve key frame indices $X^r = \{i^r_1, i^r_2, ..., i^r_{n_{rc}}\}, X^t = \{i^t_1, i^t_2, ..., i^t_{n_{tc}}\}$. n_{rc} and n_{tc} are the number of the reference key frames and that of the target ones, respectively. Unless otherwise noted, all the frames but the key frames are to be disregarded in the ensuing processes. VC is performed in a chronological manner as in Algorithm 1, where $cos\_sim(v_1, v_2)$ is the cosine similarity value of two vectors v_1, v_2. Note that this step processes independently the target video and the reference one.

Algorithm 1 Video Compression

In: S, T_c // S: set of video frames. T_c: threshold.
Out: X // X: set of key frame indices.

```
array X ← {1}
l ← 1
while l < S.length do
    for h = l + 1 to S.length do
        if cos_sim(VGG16'(S[l]), VGG16'(S[h])) < T_c then
            X.append(h)
            break
        end if
    end for
    l ← h
end while
return X
```

Proposed Temporal Alignment (TA)

TA matches frames in the target video to those in the reference video. This paper proposes a new TA method using deep features. Let $\hat{X} = \{\hat{k}_1, \hat{k}_2, ..., \hat{k}_{n_{tc}}\}$ be the indices of the matched reference frames. $\hat{X}$ is calculated by Equation (1).

$$\hat{k}_l = \arg\max_h \ cos\_sim(VGG16'(I^t_{i^t_l}), VGG16'(I^r_{i^r_h})), \tag{1}$$

where $l = 1, 2, ..., n_{tc}$. To prevent abruption from the previously matched index, a search range is restricted as $h \in \{m^{back}_l, m^{back}_l + 1, ..., m^{front}_l\}$. m^{back}_l, m^{front}_l are calculated by Equations (2), (3), respectively.

$$m^{back}_l = \begin{cases} i^r_1 & \text{if } l = 1, \\ \max(i^r_1,\ \hat{k}_{l-1} - back) & \text{otherwise,} \end{cases} \tag{2}$$

$$m^{front}_l = \begin{cases} front & \text{if } l = 1, \\ \min(i^r_{n_{rc}},\ \hat{k}_{l-1} + front) & \text{otherwise,} \end{cases} \tag{3}$$

where $back$ and $front$ are hyperparameters, discussed in the EXPERIMENTS AND RESULTS section.

Spatial Alignment (SA)

It is impossible to take all videos in an exactly consistent angle or position. Consequently, even if TA perfectly synchronizes the input videos, the matched frame pairs will have different image planes. Thus, SA, or image registration, is performed with Homography transformation. Let $H(I_1, I_2)$ be a transformed image of I_1 to I_2. The outside of an image plane is treated as black (i.e. RGB value (0, 0, 0)), when it gets into the image plane by Homography transformation. This black area would be detected as change in the following stage. Therefore, the counterpart of the target frame is masked. I' signifies a masked version of an image I.

Proposed Frame Comparison (FC)

The last component of the proposed method is FC. FC compares the spatiotemporally aligned frame pairs in the upstream processes and calculates dissimilarity maps. After that, it computes change maps by binarizing the dissimilarity maps with a threshold. Dissimilarity maps are obtained based on the idea by (Kim et al. 2017). They proposed a similarity calculation method with the VGG13 network for template matching. Feature maps of a template image and a target one were extracted from a mid-convolutional-layer of the network. The target feature map was searched with a sliding window in order to determine the patch of the target image most similar to the template image. They used Normalized Cross Correlation (NCC) as a similarity criterion.

A feature map of the m-th target frame and that of the matched reference frame are denoted as $M^t_m = VGG13((I^t_m)')$, $M^r_m = VGG13(H(I^r_{\hat{k}_m}, I^t_m))$, respectively. Note that the target frames and the reference ones, in the setting of this paper, have the same resolution, meaning their feature maps are of the same shape. This fact enables the feature maps to be compared in a position-wise manner as $SM_{m,i,j} = cos\_sim(M^t_{m,i,j}, M^r_{m,i,j})$, where $i \in \{1, 2, ..., H\}$, $j \in \{1, 2, ..., W\}$, and $M^t_{m,i,j}$ and $M^r_{m,i,j}$ are C-dimensional vectors. It is noteworthy that NCC corresponds to cosine similarity when the target pair is two vectors. By following (Kim et al. 2017), one can obtain a similarity map since it was proposed for template matching. Thus, it is converted into a dissimilarity map as $DM_{m,i,j} = 1 - SM_{m,i,j}$.

A multi-scale option is introduced as in (Carvalho et al. 2019) with some modifications. (Carvalho et al. 2019) obtained different-scale maps by resizing an frame. Subsequently, they resized them to the input size and just added them up. This way can be followed, but there are some constraints because of the CNN attribution. Some CNN layers downsample an image. While their processing, they would discard the right-end or bottom-end information of the input image not even considering it due to the filter size or the stride of those layers. VGG13 has five pooling layers, the window size of which is 2×2. The other layers of VGG13 do not affect the output size. For this reason, the resolution of the input should be divisible by $2^5 = 32$ to avoid loss of spatial information. Moreover, the aspect ratio of the proposed dataset is 16:9. Putting these conditions together, two resolution candidates are obtained: 512×288 and 1024×576. This paper extracts features only from the final pooling layer of VGG13. The feature maps from it contain the most abundant peripheral context than those from the preceding layers.

Three weight types are proposed to combine different-scale dissimilarity maps DM^k, $k \in \{1, 2, ..., n_{dm}\}$. n_{dm} denotes the total number of DM. Equation (4) shows how to create a weighted map, $weighted\_DM$.

$$weighted\_DM_{i,j} = \sum_k w^k rDM^k_{i,j}, \tag{4}$$

where rDM^k is the resized DM^k to the input size (W^{org}, H^{org}) with nearest neighbor interpolation. $i \in \{1, 2, ..., W^{org}\}$ and $j \in \{1, 2, ..., H^{org}\}$ are xy-coordinate positions. w^k is the k-th weight. One weight type is MAX as in Equation (5).

$$w^k = \frac{\max(DM^k)}{\sum_{l=1}^{n_{dm}} \max(DM^l)}. \tag{5}$$

A change is more detectable by a suited-scale map than the other scale maps. Thus, using a certain-scale map probably results in higher dissimilarity values for the corresponding-scale change than the other scale maps. Based on this idea, the MAX weight type is designed not to miss changes. Another is EQUAL as in Equation (6).

$$w^k = \frac{1}{n_{dm}}. \tag{6}$$

In EQUAL, all weights have the same value. The third weight type is LARGE as in Equation 7. Assume the

size of DM^1 be the smallest of the maps DM_1, DM_2, ..., $DM_{n_{dm}}$ and set w^1 as the possible maximum weight.

$$w^k = \begin{cases} \dfrac{1}{1 + \sum_{l=2}^{n_{dm}} \max(DM^l)} & \text{if } k = 1, \\ \dfrac{\max(DM^k)}{1 + \sum_{l=2}^{n_{dm}} \max(DM^l)} & \text{otherwise.} \end{cases} \quad (7)$$

With the LARGE weight, the smallest dissimilarity map is assigned with the possible maximum weight. In other words, the weights of the other scale maps would have smaller weights than the case of the other weight types. The smallest map contains the spatially roughest information, meaning it includes less environmental effects such as parallax than the other maps. Thus, the LARGE weight is expected to mitigate environmental effects causing false alarm. Once the weighted map is obtained by Equation (4), a change map can be calculated by binarizing the weight map. The threshold value is discussed in the EXPERIMENTS AND RESULTS section.

EXPERIMENTS AND RESULTS

Dataset

The existing datasets do not contain anomalous changes. Therefore, a new dataset has been created by recording some looming-motion videos with a radio-controlled vehicle. The vehicle ran on a straight corridor at a speed of 0.5 meters per second, and never moved backward. Its trajectories were not identical, and the viewing angle of the vehicle was inconsistent. The proposed dataset includes two sets of a reference and six target videos about 1.5 minutes long each, so the total number of the videos is fourteen. The two sets were captured at different times of the day: day and night. Table 1 shows what kind of changes the target videos contain. Each of the target videos was temporally, spatially aligned to the time-wise corresponding reference video. For TA assessment, each target frame was temporally aligned to a reference one at hand. Subsequently, dissimilarity maps were calculated by comparing each of the aligned frame pairs. To evaluate FC performance, each change was labeled with a bounding box. The proposed dataset is challenging due mainly to parallax or strong illumination change.

Parameter setting

The proposed method has some adjustable parameters. T_c was set to 0.995. For TA, $back$ was fixed to zero since the camera never moved backward in the dataset. Preferable values for *front* were roughly searched for by grid search with a set of values (3, 5, 7, 10). Consequently, this paper chose *front*= 7 for the day targets and *front*= 3 for the night ones. As referred to in the METHOD section, a multi-scale option was employed, and the input images were resized to two scales: 512×288 and 1024×576. For ZNCC, this paper followed (Carvalho et al. 2019) and prepared scales: 20×11, 40×22, 80×45, and 160×90. Besides, another scale 320×180 was also tested for a deeper survey. The window size of ZNCC was set to five.

Table 1: Change Types in the Proposed Dataset

time	data name	included change type
day	fallen	person (fallen)
		bottle
	standing	person (standing)
		umbrella (dropped)
	walking	person (walking)
		umbrella (leaning)
		door
	stacked	two boxes (stacked)
	separate	two boxes (separate)
	bag	bag
night	fallen	person (fallen)
		bottle
		shoe
	standing	person (standing)
		umbrella (leaning)
		shoe
	walking	person (walking)
		umbrella (dropped)
		shoe
	stacked	two boxes (stacked)
	separate	two boxes (separate)
	bag	bag

Experiment

The proposed TA method was compared with (Evangelidis and Bauckhage 2013). To give a quantitative comparison, this paper followed (Diego et al. 2013). They set a ground-truth interval $[l_t, u_t]$ for each index t of a sequence and calculated TA errors as in Equation (8).

$$err(t, \hat{k}_t) = \begin{cases} 0 & \text{if } l_t \le \hat{k}_t \le u_t, \\ \min(|l_t - \hat{k}_t|, |u_t - \hat{k}_t|) & \text{otherwise,} \end{cases} \quad (8)$$

where $\hat{k}_t$ is the t-th index matched by a TA method and $t \in \{1, 2, ..., n_{rc}\}$. The one-by-one ground truth GT_t, which the proposed dataset included, was expanded by one on the negative and positive sides (i.e. $l_t = \max(1,\ GT_t - 1)$, $u_t = \min(n_{rc},\ GT_t + 1)$). Table 2 shows rates of frames with equal or less than each error. TA with VGG16' provided better results than (Evangelidis and Bauckhage 2013) in all the videos. In case of $err = 0$, the error gaps are at least 8.9% (day/separate) and at most 54.5% (day/fallen). Even when comparing the $err = 0$ results by VGG16' and the $err \le 2$ results by (Evangelidis and Bauckhage 2013), most of the former results are better. Comparing the day with the night, both methods rather struggled to align the day sequences. Looking at $err = 0$, the gaps between day/Average and night/Average are 14.9% (VGG16') and 39.6% (Evangelidis and Bauckhage 2013). This is because sunlight through windows formed different shapes on the wall and floor and affected the surrounding brightness, making the day videos of the proposed dataset more challenging. This sunlight worsened (Evangelidis and Bauckhage 2013) more strongly than VGG16' because the former method using local de-

scriptors unfortunately captured the sunlight changing its form. On the other hand, such local changes were invisible for VGG16' thanks to its GAP.

With the TA results by VGG16', the FC performance of the proposed pipeline was evaluated by the area under the receiver-operator curve (AUC). VGG13 and ZNCC with the three weight types were compared as shown in Table 3. This result only exhibits the best combination of scales: [512×288, 1024×576] for VGG and [20×11, 40×22, 80×45] for ZNCC. Table 3 indicates three notable points. Firstly, VGG13 outperformed ZNCC in all of the scenarios. Secondly, both methods performed the worst on day/walking and night/stacked for each time. What deteriorated them is investigated in the following paragraph. Thirdly, the weight type provided just a marginal difference. This is because if just a single pixel in one of the different-scale maps has a high value, it pushes up the weight of that map. Therefore, all weights ended up getting almost the same value.

Table 4 shows AUC scores of the FC methods with and without the proposed TA results. Only the LARGE weight type was shown as it performed the best. At day/walking and night/stacked in Table 4, large gaps can be seen. This suggests the failure of TA led to the terrible FC performance. This suggestion was confirmed by counting detectable pixels, which were non-masked ones in the METHOD section. The inner rate columns in Table 4 show each detectable pixel rate (%). Non-change areas are the outside of bounding boxes, and change areas are the inside. As one can see, the change inner rates for day/walking and night/stacked are obviously low, meaning a large part of the change areas was regarded as unchanging. Therefore, Table 4 proves TA plays a pivotal role in CD.

This paper expanded on how scale sizes affected results. Table 5 shows AUC scores for VGG13 and ZNCC with single scales. Smaller resolutions provided better scores because the AUC was a pixel-wise criterion. That is, a method tuned for larger objects contributes to the score more than smaller ones. Also, this tendency can be found in the best set of scales for ZNCC ([20×11, 40×22, 80×45]). Another notable point is that, comparing the AUC scores in Table 5 with the AUC scores using the TA ground truth in Table 4, the combination of multi scales enhanced the detection ability. In addition to the AUC, this paper looked into the relationship of scales and dissimilarity values for each object. This paper calculated the median of dissimilarity values belonging to each change type or the background and then a ratio of each object median to the background one. If a ratio is less than 1.0, the corresponding change is indistinguishable from the background. The higher it is, the more detectable the change is. Note that the looming motion in the videos significantly varies the size of changes. Thus, Table 6 only shows "person" and "bottle", a large and a relatively small change, in {day, night}/fallen. One can see the tendency of larger resolutions spotting smaller changes and vice versa.

Finally, to discuss the FC performance for each change type, a ratio of each object median to the background one was computed with VGG13 ([512×288, 1024×576]) and ZNCC ([20×11, 40×22, 80×45]) as shown in Table 7. The weight type was fixed to LARGE as in Table 4. Table 7 indicates some characteristics of ZNCC and VGG13. ZNCC shows distinctively strong and weak points. It failed to detect the smallest change, "bottle". Moreover, the value for a relatively small object "umbrella" is significantly smaller than the other changes except for "door". Note that although "shoe" might sound small, it appears close to the vehicle trajectories. Thus, "shoe" looks big in the proposed dataset. On the other hand, VGG13 successfully detected "bottle". Its ratio is actually close to 1.0, but this result seems reasonable because "bottle" is not only small but also unobtrusive in the proposed dataset. For the other changes including "umbrella", VGG13 almost impartially spotted them. This implies that VGG13 does not largely depend on the input scales. This is because its convolutional layers acquire surrounding information.

CONCLUSION

This paper aims to automatically monitor areas for security using a moving camera instead of humans. None of the existing CD datasets was designed for such a purpose. Thus a new dataset for area surveillance has been built with a UGV. Subsequently, this paper has introduced a structured method and devised three components for it: VC, TA, and FC. For FC, three ways to combine different-scale maps have also been proposed. To perform an evaluation, the proposed TA and CD methods were compared with classic methods. Through the experiments, this paper showed the effectiveness of the proposed method in area surveillance using a moving camera.

There are some limitations in the proposed method. First, the proposed CD method cannot detect changes in a target frame if the matched reference frame does not contain the spatially corresponding region. In terms of false positive, there were some times the method falsely detected objects as changes due to difference in viewing angle or position and the sunlight. Second, the FC performance strongly depends on the preceding procedure: TA and SA, as shown in Table 4. Third, if changes appear in a dominant part of an image, TA and SA would provide a poor result. Finally, the reference video has to contain the whole scenes of the target video. This limits a range of applications.

A piece of the future work is to improve the proposed method by overcoming the limitation. It is necessary to research how to make methods robust to environments. Evaluation-wise, this paper performed an evaluation with the pixel-wise AUC. As aforementioned, it tended to give better scores to a method tuned for larger changes. This tendency is not appropriate for surveillance. For this reason, a new frame-level evaluation should be consid-

Table 2: Frame Rates (%) with Equal or Less than Each Error for TA

		VGG16'			georgios		
time	data	$err = 0$	$err \leq 1$	$err \leq 2$	$err = 0$	$err \leq 1$	$err \leq 2$
day	fallen	68.1	80.1	84.3	13.6	20.4	26.7
	standing	52.3	70.5	76.7	8.0	11.9	14.8
	walking	55.7	73.4	83.7	18.2	26.1	34.5
	stacked	78.7	88.1	94.1	38.6	46.0	49.0
	separate	78.8	90.2	90.7	69.9	85.0	89.1
	bag	67.7	80.8	83.8	18.2	23.7	27.8
	Average	66.9	80.5	85.6	27.8	35.5	40.3
night	fallen	91.3	96.9	100.0	81.6	87.2	91.3
	standing	95.3	97.9	98.4	68.6	79.1	84.3
	walking	98.5	100.0	100.0	85.8	92.9	93.9
	stacked	79.7	91.9	97.7	54.7	65.7	70.9
	separate	68.0	74.6	80.1	27.6	40.3	51.9
	bag	97.8	100.0	100.0	86.0	91.6	98.3
	Average	81.8	93.6	96.0	67.4	76.1	81.8

Table 3: AUC Scores with the Proposed TA Component for FC

		VGG13			ZNCC		
time	data	MAX	EQUAL	LARGE	MAX	EQUAL	LARGE
day	fallen	0.870	**0.871**	**0.871**	0.745	0.739	0.738
	standing	**0.842**	**0.842**	**0.842**	0.731	0.730	0.729
	walking	0.686	0.688	**0.689**	0.654	0.661	0.663
	stacked	0.860	0.862	**0.863**	0.708	0.709	0.704
	separate	0.813	0.817	**0.818**	0.765	0.784	0.786
	bag	**0.926**	0.925	0.925	0.812	0.803	0.801
	overall	0.833	0.834	**0.835**	0.736	0.738	0.737
night	fallen	0.913	0.914	**0.914**	0.834	0.837	0.838
	standing	0.915	**0.916**	**0.916**	0.821	0.820	0.820
	walking	0.917	0.921	**0.922**	0.864	0.877	0.879
	stacked	**0.584**	0.583	0.583	0.512	0.513	0.515
	separate	**0.899**	**0.899**	**0.899**	0.837	0.840	0.841
	bag	0.901	0.903	**0.904**	0.808	0.812	0.811
	overall	0.855	**0.856**	**0.856**	0.779	0.783	0.784
Average		0.844	**0.845**	**0.845**	0.758	0.760	0.760

Table 4: Change/Non-Change Inner Rates (%) and AUC Scores with and without the Proposed TA Component

		TA_VGG16'				TA_GROUND_TRUTH			
time	data	non-change inner rate	change inner rate	VGG13	ZNCC	non-change inner rate	change inner rate	VGG13	ZNCC
day	fallen	83.2	97.1	0.871	0.738	86.9	98.6	**0.908**	0.770
	standing	80.2	92.9	0.842	0.729	85.4	90.8	**0.869**	0.754
	walking	79.4	69.8	0.689	0.663	82.0	85.9	**0.831**	0.738
	stacked	86.7	100.0	**0.863**	0.704	88.3	100.0	0.853	0.691
	separate	92.3	86.2	**0.818**	0.786	93.6	81.1	0.798	0.786
	bag	81.4	97.9	0.925	0.801	83.5	100.0	**0.966**	0.862
night	fallen	92.6	98.5	**0.914**	0.838	92.5	97.9	**0.914**	0.842
	standing	87.9	99.6	0.916	0.820	88.1	98.1	**0.929**	0.841
	walking	91.2	95.1	0.922	0.879	90.5	97.3	**0.936**	0.889
	stacked	86.7	64.9	0.583	0.515	88.5	98.1	**0.861**	0.727
	separate	82.5	98.6	0.899	0.841	87.3	99.4	**0.940**	0.869
	bag	93.5	93.4	0.904	0.811	94.2	93.4	**0.911**	0.815
Average				0.845	0.760			**0.893**	0.799

Table 5: AUC Scores for Different Scales with the TA Ground Truth

		VGG13		ZNCC				
time	data	512×288	1024×576	20×11	40×22	80×45	160×90	320×180
day	fallen	0.910	0.880	0.703	0.766	0.764	0.738	0.699
	standing	0.865	0.841	0.706	0.741	0.719	0.683	0.631
	walking	0.848	0.790	0.731	0.725	0.678	0.618	0.571
	stacked	0.837	0.831	0.572	0.689	0.717	0.719	0.697
	separate	0.815	0.757	0.834	0.758	0.677	0.607	0.550
	bag	0.952	0.949	0.762	0.842	0.863	0.824	0.754
	overall	0.871	0.841	0.718	0.753	0.736	0.698	0.650
night	fallen	0.908	0.890	0.807	0.830	0.818	0.768	0.710
	standing	0.919	0.905	0.787	0.835	0.804	0.704	0.609
	walking	0.936	0.895	0.891	0.865	0.819	0.740	0.663
	stacked	0.833	0.854	0.686	0.709	0.726	0.728	0.699
	separate	0.936	0.918	0.847	0.856	0.834	0.796	0.747
	bag	0.909	0.880	0.793	0.815	0.793	0.707	0.599
	overall	0.907	0.890	0.802	0.818	0.799	0.740	0.671
Average		0.889	0.866	0.760	0.786	0.768	0.719	0.661

Table 6: The Ratio of Each Median of Two Objects to the Background One for Each Scale

		VGG13		ZNCC				
time	change	512×288	1024×576	20×11	40×22	80×45	160×90	320×180
day	person	3.39	2.76	1.90	4.36	4.80	3.58	2.54
	bottle	2.78	2.64	1.20	1.09	3.40	2.54	2.14
night	person	4.95	3.13	13.50	20.00	13.67	5.41	2.64
	bottle	0.92	1.64	0.50	0.33	0.33	1.29	1.15

Table 7: The Ratio of Each Object Median to the Background One

	person	bottle	umbrella	door	box	bag	shoe
VGG13	3.23	1.46	2.82	2.80	3.30	3.38	2.91
ZNCC	6.10	0.70	2.70	3.20	6.40	5.50	4.40

ered. Finally, the proposed dataset contains little variation. Thus, it is required to record videos in different seasons or weather. On top of that, other places such as curves should be included.

REFERENCES

Alcantarilla, P. F.; S. Stent; G. Ros; R. Arroyo; and R. Gherardi. 2018. "Street-view change detection with deconvolutional networks", *Journal of Autonomous Robots*, Vol. 42 (May), 1301-1322.

Carvalho, G. H. F. de; L. A. Thomaz; A. F. da Silva; E. A. B. da Silva; and S. L. Netto. 2019. "Anomaly Detection with a Moving Camera using Multiscale Video Analysis", *Journal of Multidimensional Systems and Signal Processing*, Vol. 30, Issue 1 (January), 311-342.

Chu, Wenqing; H. Xue; C. Yao; and D. Cai. 2019. "Sparse Coding Guided Spatiotemporal Feature Learning for Abnormal Event Detection in Large Videos", *IEEE Transactions on Multimedia*, Vol. 21, Issue 1, 246-255.

Deng, J.; W. Dong; R. Socher; L.-J. Li; K. Li; and L. Fei-Fei. 2009. "ImageNet: A large-scale hierarchical image database", *Proceedings of* 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, 20-25.

Diego, F.; J. Serrat; and A. M. López. 2013. "Joint Spatio-Temporal Alignment of Sequences", *Journal of IEEE Transactions on Multimedia*, Vol. 15, No. 6 (October), 1377-1387.

Evangelidis, G. D. and C. Bauckhage. 2013. "Efficient Subframe Video Alignment using Short Descriptors", *Journal of IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 10 (October), 2371-2386.

Hao, Y.; Z.-J. Xu; Y. Liu; J. Wang; and J.-L. Fan. 2019. "Effective Crowd Anomaly Detection through Spatio-temporal Texture Analysis", *Journal of Automation and Computing*, Vol. 16, Issue 1 (February), 27-39.

Kim, J.; J. Kim; S. Choi; M. A. Hasa; and C. Kim. 2017. "Robust Template Matching using Scale-Adaptive Deep Convolutional Features", *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 708-711.

Lin, M.; Q. Chen; and S. Yan. 2014. "Network In Network", *Proceedings of International Conference on Learning Representations*, 10 pages.

Morais R.; V. Le; T. Tran; B. Saha; M. Mansour; and S. Venkatesh. 2019. "Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 11996-12004.

Sakurada, K. and T. Okatani. 2015. "Change detection from a street image pair using CNN features and superpixel segmentation", *Proceedings of British Machine Vision Conference*, 12 pages.

Silva, A. F. da; L. A. Thomaz; G. Carvalho; M. T. Nakahata; E. Jardim; J. F. L. de Oliveira; E. A. B. da Silva; S. L. Netto; G. Freitas; and R. R. Costa. 2014. "An Annotated Video Database for Abandoned-Object Detection in a Cluttered Environment", *Proceedings of* 2014 *International Telecommunications Synposium*, 5 pages.

Simonyan, K. and A. Zisserman. 2014. "Very deep convolutional networks for large-scale image recognition", *Preceedings of International Conference on Learning Representations*, 14 pages.

Singh, A.; D. Patil; and S. N. Omkar. 2018. "Eye in the Sky: Real-Time Drone Surveillance System (DSS) for Violent Individuals Identification Using ScatterNet Hybrid Deep Learning Network", *Proceedings of* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), 1710-1718.