UNUSUAL EVENTS DETECTION BASED ON MULTI-DICTIONARY SPARSE REPRESENTATION USING KINECT

Can Wang, Hong Liu

Key Laboratory of Machine Perception and Intelligence
Shenzhen Graduate School, Peking University, 100871, Beijing, China
E-mail : {canwang, hongliu}@pku.edu.cn

ABSTRACT

Unusual events detection plays a crucial role in surveillance applications, which is becoming more and more urgent need for public security. However, illumination and scale changing, lacking of sufficient training data and subjective of abnormality definition are some of the severe difficulties, which are hard to deal with by widely used traditional cameras. In order to solve these problems, first, a novel feature is proposed in this paper, which is named random local feature (RLF) to describe the spatial-temporal information of depth image detected by the Kinect sensor. Then, we expand the sparse representation framework to a multi-dictionary sparse representation framework, based on the intuition that that anomaly of a same event may vary a lot in different regions in a scene. We split the depth video into several regions and use detected RLF features in each region to train dictionary by K-SVD algorithm, and use the OMP algorithm to sparse-represent each feature. Finally, an objective function is introduced to evaluate the anomaly of features in each region according to reconstruction errors. Unusual events are defined as those incidences that occur very rarely in the entire video sequence in our system, which is tested on real data and demonstrates promising results in unusual events detection.

Index Terms— Kinect, Sparse Representation, Anomaly Detection

1. INTRODUCTION

Anomaly detection is an active area of research over these years, and an in-depth review of its literature can be found in a recent survey by Chandola et al. [1][2]. The major difficulties in video-based anomaly detection are definition of abnormal events, how to describe events. Moreover, as is often the case, one of the major difficulties in video analysis is the huge amount of data [3].

In traditional abnormal events detection framework, most systems are supervised [4, 5, 6] and model-based. Researchers should define what is abnormal event, but the definition is rather subjective. While it is often true that only a small portion of video contains abnormal events while all others are normal. So it is hard to training classifiers for abnormal events. Another category is clustering-based [7, 8, 9, 10], but they also need prior assumption on what anomaly is. As anomaly definition is subjective and even vary a lot in different scenes, this category is not adopted in our framework.

In this work, we adopt a recently rising category using sparse coding framework [3] which is used to detect unusual events in videos. In this framework, usual events are more likely to be reconstructible from an event dictionary while unusual events are not, and the definition of normality and abnormality is changed online. It is built upon a rigorous statistical principle, and makes no prior assumptions of what unusual events may look like, hence no need to obtain prior models, templates, knowledge of the clusters and is completely unsupervised only based on the assumption that an unusual event is unlikely to occur in the small initial portion of a video. In order to difficulties of events description based on RGB video, such as perspective effects (scale problem), vulnerable to bad illumination conditions, similarity between foreground and background, etc., we originally only use depth data for events description for two reasons: First, depth data is invariant to scale , color, texture and illumination changing. Second, recently depth sensor like Kinect sensor [11] is inexpensive and can fast capture dense depth map of the scene. Third, depth video essentially reflects motion of the scene which makes it suitable for events description.

Our main contribution lies in two aspects: First, a multi-dictionary sparse representation framework is proposed for unusual events detection using depth video detected by Kinect, and a brief illustration is shown in Fig.1. Unusual events are defined as those incidences that occur very rarely in the entire video sequence [3, 12, 13]. Second, in the events representation module, a novel feature (RLF feature) is designed to describe the local spatio-temporal information surrounding forward motion salient points(FMSP), which are extracted from forward motion regions. Based on FMSP salient points, RLF features are able to better describe motion information in the scene. Detailed formulation of RLF feature is described in Section 2. Based on RLF features, the K-SVD algorithm [14] is adopted for multi-dictionary training, which is suitable for efficient multi-thread programming. And the OMP algorithm [15] is used to reconstruct RLF features. The reconstruction errors are used to evaluate the anomaly. The algorithm is tested in several real indoors scenes and experimental results verify its effectiveness and efficiency.

2. DEPTH VIDEO REPRESENTATION

Our motivation is to find a better feature to describe depth video. First we choose the 3D-SIFT feature [16] and a spatio-temporal invariant feature STIP [17] which outperforms the current state-of-the art in grayscale image processing. But for the depth image, it's difficult get acceptable results. Different from grayscale or RGB image, a pixel in a depth image essentially represents a 3D position in the world coordinates and a depth video essentially represents the variation of 'the position', say, motion. So a motion-based salient feature is the first choice to try for depth video representation.

2.1. Forward Motion Salient Points

In our depth image sequences, a lower pixel value indicates a larger range from the Kinect sensor. First, we propose a kind of motion salient point: forward motion salient point (FMSP), formulated as

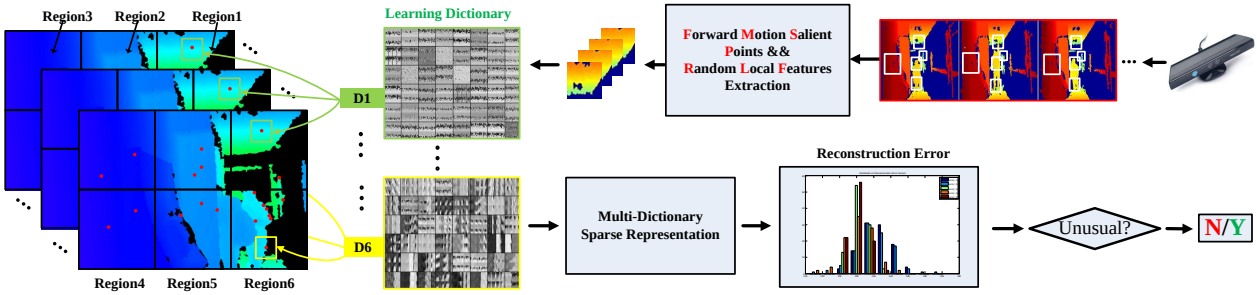


Fig. 1: Overview of our proposed method. Left part is a sparse representation module where FMSP points (red) are detected in different regions and RLF descriptors (cuboids along time axis) are used for training multi-dictionary. Right part is simple illustration of unusual events detection: its top part is a process of feature extraction; its bottom part is a process of features reconstruction process where the reconstruction error is used for unusual evaluation.

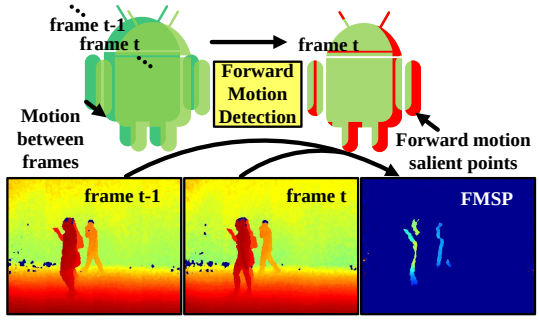


Fig. 2: Illustration of Forward Motion salient points

below:

$$\begin{aligned} & \boldsymbol{X}_{fm} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} = \{\boldsymbol{x}_i\}_{i=1}^{n} \\ s.t. \quad & \forall \boldsymbol{x}_i \in \boldsymbol{X}_{fm}, \quad I_t(\boldsymbol{x}_i) - I_{t-1}(\boldsymbol{x}_i) > \tau_{fm} \\ & I_{t-1}(\boldsymbol{x}_i) > \tau_{ur} \end{aligned} \tag{1}$$

At a given image coordinate $\boldsymbol{x}_i$, $I_t(\boldsymbol{x}_i)$ is the pixel value of depth image I at time t. τ_{fm} is a threshold indicating whether there is a significant range change in $\boldsymbol{x}_i$. τ_{ur} is set in order to avoid influence of unstable range change [11] commonly in depth video. For an intuitively understanding of the forward motion (FM), an illustration is given in Fig.2.

In practical applications, raw depth data always contains considerable number of regions with unstable pixel value, like the irregular empty holes shown the first two images in bottom row in Fig.2. This is normally caused by hardware drawbacks of the Kinect sensor [11], which always brings wrong detection of FMSP salient points. In order to reduce wrong salient points in salient points set and reduce its redundancy, m points are randomly selected from $\boldsymbol{X}_{fm}$ and normally $m \ll n$. Assume there are N points in $\boldsymbol{X}_{fm}$:

$$\begin{aligned} & \boldsymbol{X}_{rfm} = \{\boldsymbol{x}_1^r, ..., \boldsymbol{x}_m^r\} = \{\boldsymbol{x}_i^r\}_{i=1}^{m} \\ s.t. \quad & \forall \boldsymbol{x}_i^r \in \boldsymbol{X}_{fm},\ \boldsymbol{x}_i^r \ is\ selected\ randmly \end{aligned} \tag{2}$$

Normally, the unstable region points being randomly selected have low possibility up to a certain amount $\tau_m m$:

$$\begin{aligned} & P\{m_u > \tau_m m\} < P\{m_u = \tau_m m\} = \frac{C_{N_{ur}}^{\tau \cdot m} \cdot C_{N_{fm}}^{(1-\tau_m)m}}{C_N^m} \\ & \approx \frac{(\tau_u N)^{\tau_m \cdot m} \cdot [(1-\tau_u)N]^{(1-\tau_m)m}}{N^m} < 0.01 \\ & s.t. \quad \tau_m \approx 0.1,\ \tau_u \approx 0.05,\ m \ll N \end{aligned} \tag{3}$$

where τ_u is the ratio of unstable region points in all N salient points in the original set $\boldsymbol{X}_{fm}$. However, only motion salient points cannot represent global status of the depth video. For describing other parts of the video, the final salient points set $\boldsymbol{X}$ contains 25% points $\boldsymbol{X}_{gl}$ which are randomly selected from global depth video as well as contains 75% FMSP points $\boldsymbol{X}_{rfm}$ to emphasize the motion-salient parts:

$$\begin{aligned} & \boldsymbol{X} = \{\boldsymbol{X}_{gl}, \boldsymbol{X}_{rfm}\} \\ s.t. \quad & \boldsymbol{card}(\boldsymbol{X}_{gl}) = \lambda_0 \boldsymbol{card}(\boldsymbol{X}_{rfm}),\ \lambda_0 = 0.2 \end{aligned} \tag{4}$$

2.2. Random Local Feature

Inspired by depth image features proposed in [18], we design a random local feature (RLF) to describe the local spatio-temporal information around the salient point. Each feature is a R dimension vector with each element randomly formulated in a local spatio-temporal cuboid as following:

$$\begin{aligned} & f_\theta(\boldsymbol{\Pi}, \boldsymbol{x}, t) = I_{t'}\left(\boldsymbol{x} + \frac{\boldsymbol{u}}{d_{I_{t'}}(\boldsymbol{x})}\right) - I_{t'}\left(\boldsymbol{x} + \frac{\boldsymbol{v}}{d_{I_{t'}}(\boldsymbol{x})}\right) \\ s.t. \quad & \theta = (\boldsymbol{u}, \boldsymbol{v}, \Delta t), \quad \boldsymbol{\Pi} = \{I_{t-\tau_c}, ..., I_t\}, \\ & \Delta t \in [0, \tau_c], \quad \boldsymbol{x} \in \boldsymbol{X}, \quad t' = t - \Delta t \\ & d_{I_{t'}}(\boldsymbol{x}) = \lambda_d \frac{(255 - I_{t'}(\boldsymbol{x}))}{100} + \lambda_r \end{aligned} \tag{5}$$

$$\boldsymbol{y}_i = \{f_{\boldsymbol{\theta}_j}(\boldsymbol{x}_i)\}_{j=1}^{R} = \{f_{\boldsymbol{\theta}_1}(\boldsymbol{x}_i), ..., f_{\boldsymbol{\theta}_R}(\boldsymbol{x}_i)\} \tag{6}$$

where $d_{I_{t'}}(\boldsymbol{x})$ indicates the real world range (depth) at $\boldsymbol{x}$ in depth image $\boldsymbol{I}$, and parameter $\boldsymbol{\theta} = (\boldsymbol{u}, \boldsymbol{v}, \Delta t)$ contains offsets $\boldsymbol{u}$, $\boldsymbol{v}$ and Δt in image coordinates and time axis respectively. The normalization of the offsets $\boldsymbol{u}$ and $\boldsymbol{v}$ by $\frac{1}{d_{I_{t'}}(\boldsymbol{x})}$ ensures the features are depth invariant: at a given point in depth image, a fixed world space offset will result whether the pixel is close or far from the camera [18]. And $\boldsymbol{u}$, λ_d and λ_r depend on parameters of Kinect sensor. Supposing there are N salient points in $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, the depth video can be represented as RLF feature descriptors based on all salient points:

$$\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{N} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N\} \tag{7}$$

3. MULTI-DICTIONARY SPARSE REPRESENTATION FOR UNUSUAL EVENTS DETECTION

In previous work [19][3][20], no matter what statistical models were adopted for unusual events detection, usually use one statistic model

to represent the global video. Actually they have a common implicit assumption that abnormality criterion is unique in the global video, however, this is not the case in many environments. For example, walk on road and walk on roof show totally different abnormality. In our framework, the image coordinates are split into M regions averagely, formulated as $\boldsymbol{B} = \{\boldsymbol{b}_1, ..., \boldsymbol{b}_M\}$, shown in Fig. 1. According to this division $\boldsymbol{B}$, RLF features set $\boldsymbol{Y}$ is also split into M parts, formulated as below:

$$\begin{aligned} &\boldsymbol{Y} = \{\boldsymbol{Y}_k\}_{k=1}^{M} = \{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_M\} \\ s.t. \quad &\boldsymbol{Y}_k = \{\boldsymbol{y}_k^1, ..., \boldsymbol{y}_k^{N_k}\}, \quad \sum_{k=1}^{M} N_k = N \end{aligned} \tag{8}$$

K-SVD Algorithm for Dictionary Training: K-SVD has two modes of operation: sparsity-based and error-based. For sparsity-based minimization, the optimization problem is given by:

$$\min_{\boldsymbol{D},\boldsymbol{\Gamma}} \|\boldsymbol{Y} - \boldsymbol{D} * \boldsymbol{\Gamma}\|_F^2 \quad s.t. \, \|\boldsymbol{\gamma}_i\|_0 \leq T, \, \boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_N] \tag{9}$$

where $\boldsymbol{Y}$ is the set of training signals, $\boldsymbol{\gamma}_i$ is the i_{th} column of $\boldsymbol{\Gamma}$, and T is the target sparsity. For error-based minimization, the optimization problem is given by:

$$\min_{\boldsymbol{D},\boldsymbol{\Gamma}} \|\boldsymbol{\Gamma}\|_0 \quad s.t. \, \|\boldsymbol{y}_i - \boldsymbol{D} * \boldsymbol{\gamma}_i\|_2 \leq \epsilon \tag{10}$$

where $\boldsymbol{y}_i$ is the i_{th} training signal, and ϵ is the target error.

Multiple Dictionaries Formulation: For training signals in each video region block $\boldsymbol{b}_k$, its dictionary $\boldsymbol{D}_k$ can be formulated using K-SVD algorithm:

$$\boldsymbol{D} = \{\boldsymbol{D}_k\}_{k=1}^{M} = \{\boldsymbol{D}_1, ..., \boldsymbol{D}_M\} \tag{11}$$

Sparsity-based approach is adopted in our framework, and a sub-dictionary $\boldsymbol{D}_k$ is formulated as following:

$$\begin{aligned} &\min_{\boldsymbol{D}_k,\boldsymbol{\Gamma}_k} \|\boldsymbol{Y}_k - \boldsymbol{D}_k * \boldsymbol{\Gamma}_k\|_F^2 \\ s.t. \quad &\left\|\boldsymbol{\gamma}_k^i\right\|_0 \leq T, \quad \boldsymbol{\Gamma}_k = [\boldsymbol{\gamma}_k^1, ..., \boldsymbol{\gamma}_k^{N_k}] \end{aligned} \tag{12}$$

OMP Algorithm for Sparse Representation: OMP is short for Sparsity-constrained Orthogonal Matching Pursuit. Given the trained dictionary $\boldsymbol{D}_k$ and the input signal (here is RLF feature) $\boldsymbol{y}_k$, it solves the optimization problem formulated as following:

$$\min_{\boldsymbol{\gamma}_k} \|\boldsymbol{y}_k - \boldsymbol{D}_k * \boldsymbol{\gamma}_k\|_2 \quad s.t. \quad \|\boldsymbol{\gamma}_k\|_0 \leq T \tag{13}$$

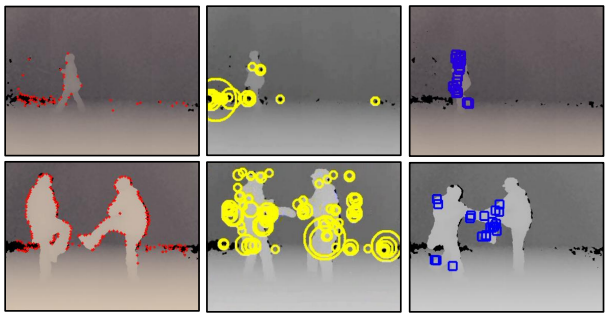
In our framework, given trained $\boldsymbol{D}_k$ and input signal $\boldsymbol{y}_k$, we adopt OMP algorithm to get its sparse representation $\boldsymbol{\gamma}_k$.

Unusual Events Detection: Given the trained sub-dictionary D_k, the input RLF feature $\boldsymbol{y}_k$ located in kth region, and its sparse representation $\boldsymbol{\gamma}_k$, the following objective function is adopted to measure the abnormality of $\boldsymbol{y}_k$:

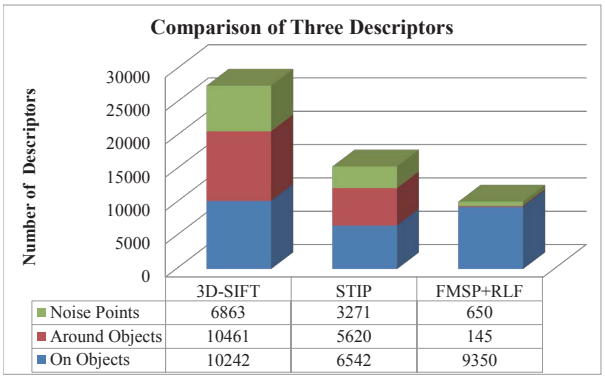
$$\mathrm{E}(\boldsymbol{y}_k, \boldsymbol{\gamma}_k, \boldsymbol{D}_k) = \|\boldsymbol{y}_k - \boldsymbol{D}_k * \boldsymbol{\gamma}_k\|_2 \tag{14}$$

which gives the reconstruction error between the input signal $\boldsymbol{y}_k$ and its sparse representation $\boldsymbol{\gamma}_k$. The RLF feature $\boldsymbol{y}_k$ actually describes a local spatio-temporal cuboid around a motion-salient point. $\boldsymbol{y}_k$ is detected as unusual if the following criterion is satisfied:

$$\mathrm{E}(\boldsymbol{y}_k, \boldsymbol{\gamma}_k, \boldsymbol{D}_k) > \hat{\epsilon}_k \tag{15}$$



(a) 3D-SIFT; STIP; FSMP+RLF



(b) A statistical distribution of three descriptors.

Fig. 3: Comparison between three features in describing depth video, which shows two advantages of FSMP+RLF: (1) high accuracy: most features are exactly located on moving objects. (2) less redundancy

where $\hat{\epsilon}$ is a user defined threshold that controls the sensitivity of the detection algorithm to unusual events. In our system, $\hat{\epsilon}$ is obtained by a statistical process during dictionary training, formulated as:

$$\hat{\epsilon}_k = \lambda_{\bar{\epsilon}} \cdot \frac{1}{N_t} \sum_{i=1}^{N_t} \mathrm{E}(\boldsymbol{y}_i, \boldsymbol{\gamma}_i, \boldsymbol{D}_i) \tag{16}$$

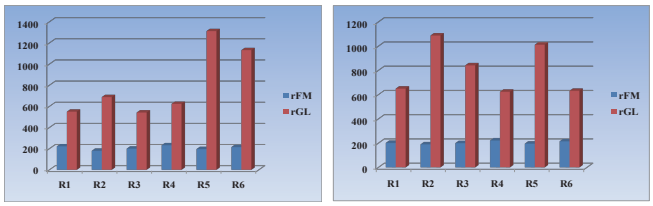
where N_t is the number of all RLF features for dictionary training.

4. EXPERIMENTS AND DISCUSSIONS

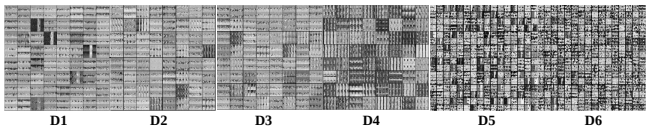
Kinect: As introduced in [11], Kinect sensor outputs video at a frame rate of 30 Hz. The RGB video stream uses 8-bit VGA resolution (640*480 pixels) with a Bayer color filter, while the monochrome depth sensing video stream is in VGA resolution (640*480 pixels) with 11-bit depth, which provides 2,048 levels of sensitivity. The Kinect sensor has a practical ranging limit of 1.2-3.5 m (3.9-11 ft) distance when used with the Xbox software, although the sensor can maintain tracking through an extended range of approximately 0.7-6 m (2.3-20 ft).

Database: Our dataset is obtained in some open environments, which includes 12 videos in a corridor of a teaching building in our university, and 6 videos in a ATM machine room of a bank[1]. Each

[1]It's a pity that we are not allowed to record videos in the subway station due to related laws and regulations, so we appreciate some courtesy from other colleagues.



(a)Corridor of a teaching building (b)ATM room of a bank



(c)Trained multi-dictionary of video in (a)

Fig. 4: Different distribution of salient points over multiple regions, which shows the necessity of multi-dictionary scheme. Each block is a base of a dictionary. First four regions have 64 bases and left two have 256 bases.

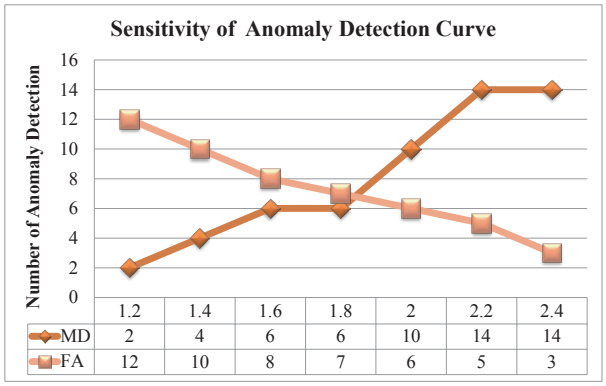


Fig. 5: As $\lambda_{\hat{\epsilon}}$ increases, the sensitivity of detection algorithm decreases, so number of MD (unusual events are not detected) increases and number of FA (usual events are detected as unusual) decreases.

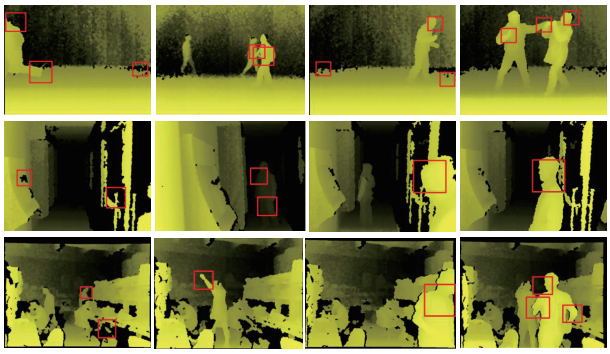


Fig. 6: Sampled qualitative experimental results in several scenes.

video lasts for around 5-15 minutes and there are 389,732 frames all together. The dataset include RGB image sequences and depth image sequences. The resolution of these images is 640*480 pixels. In our experiments, only depth image sequences are used to process.

Experiment One: A comparison experiment is conducted between RLF features (based on FMSP salient points), STIP invariant points [17] and 3D-SIFT [16], to evaluate their performance on describing salient motion and events in depth video analysis. As shown in Fig.3 (a), SURF features are most located on where there is a significant range change. Many STIP features are detected where there are significant motion, but there are also lots of STIP features detected on unstable regions. The detected FMSP salient points are all located on motion-salient objects, which makes RLF features based on them be able to better describe the depth video. In our system, $\tau_{fm} = 20$, $\tau_{ur} = 10$ in Eq.(1) and $\boldsymbol{u}$ is randomly selected in a block with side length of 10 pixels, $\lambda_d = 2m$, $\lambda_r = 0.5m$ in Eq.(5). Fig.3 (b) gives a statistical illustration of distribution of features location on object, around objects or on unstable regions.

Experiment Two: Experiments are conducted in *eight* video segments, *four* are recorded in corridor of a teaching building and four are recorded in ATM room of a bank. There are *two* usual video segments for each scene. Firstly, we trained multi-dictionary use usual video segments, and then detect unusual events in other video segments. In our system, final feature points include 80% FSMP points and 20% randomly selected points from global image. Fig.4 (a) and (b) gives the distribution histogram of these two feature points, which shows different regions have different motion saliency. Fig.4 (c) gives the visualization of the trained dictionaries in scene (a).

Experiment Three: In our framework, $\lambda_{\hat{\epsilon}}$ in Eq.(16) is a user defined threshold that is vital in unusual events detection. This experiment is conducted in order to obtain a optimal $\lambda_{\hat{\epsilon}}$ to better control the sensitivity of the detection algorithm of unusual events. An event is detected as unusual if there are one or more RLF features are satisfied Eq.(15). There are two sets of ground truth video segments: usual segments and unusual segments. 20 usual segments are sampled from usual videos for multi-dictionary training and 20 unusual segments are sampled from other depth videos. Experiments result is shown in Fig.5. The curves of missing detection (MD) and false alarm (FA) indicates $\lambda_{\hat{\epsilon}} = 1.8$ is the most optimal choice for the unusual events detection algorithm.

Finally, qualitative experimental results are given in Fig.6, sampled anomaly detection results in several scenes. The red blocks indicate abnormal motion patterns. The different size of blocks exhibit the depth-invariant features of RLF descriptors: Low-depth block gets larger size and high-depth block gets larger size. Although the false alarm is at a high level in some environments for noises or a unreasonable control of sensitivity of anomaly criterion, the system exhibits a certain level of intelligence and effectiveness on anomaly detection.

5. CONCLUSIONS

In this paper, we design a random local feature (RLF) based on forward motion salient points (FMSP) for depth video analysis. The FMSP points are usually located on motion-salient objects and the RLF feature describes local spatio-temporal information surrounding a salient point in depth video. Compared with the state-of-art features widely used in video analysis, FMSP points and RLF features can better describe the motion information of depth video. Based on these local features, we expand the sparse representation framework to a multi-dictionary sparse representation framework, based on the intuition that amomaly of a same event may vary a lot in different regions in a scene. In our future work, online update scheme for dictionary will introduced and multi-regions will split according to RLF features distribution in the scene.

6. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2010. 1

[2] Oluwatoyin P. Popoola and Kejun Wang, "Video-based abnormal human behavior recognitionła review," 2012. 1

[3] Z. Bin, F.F. Li, and E.P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *CVPR*, 2011. 1, 2

[4] X. Ma X. Wang and E. Grimson., "Unsupervised activity perception by hierarchical bayesian models.," in *CVPR*, 2007. 1

[5] S. Gong J. Li and T. Xiang., "Global behaviour inference using probabilistic latent semantic analysis.," 2008. 1

[6] J. Kim and K. Grauman., "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates.," 2009. 1

[7] A. Agovic, A. Banerjee, A. R. Ganguly, and V. Protopopescu, "Anomaly detection in transportation corridors using manifold embedding," *International Workshop on Knowledge Discovery from Sensor Data.*, 2007. 1

[8] O. Boiman and M. Irani., "Detecting irregularities in images and in video.," 2005. 1

[9] S. Batta A. Bobick C. Isbell R. Hamid, A. Johnson and G. Coleman., "Detection and explanation of anomalous activities: Representing activities as bags of event n-grams.," 2005. 1

[10] and M.Visontai. H.Zhong, J.Shi, "Detectingunusual activity in video.," 2004. 1

[11] Microsoft Corp. and Redmond WA., "Kinect for xbox 360," . 1, 2, 3

[12] J. Li, S. Gong, and T. Xiang., "Global behaviour inference using probabilistic latent semantic analysis.," in *BMVC*, 2008. 1

[13] A. B. A. Gritai and M. Shah., "Learning object motion patterns for anomaly detection and improved object detection.," in *CVPR*, 2008. 1

[14] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, 2006. 1

[15] M. Elad, R. Rubinstein, and M. Zibulevsky, "Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit," *Technical Report - CS, Technion*, 2008. 1

[16] Mubarak Shah. Paul Scovanner, Saad Ali, "A 3-dimensional sift descriptor and its application to action recognition.," *ICMM*, 2007. 1, 4

[17] I.Laptev, "On space-time interest points," *IJCV*, 2005. 1, 4

[18] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011. 2

[19] K. Yan, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in *ICCV*, 2007. 2

[20] X. Song, X. Shao, R. Shibasaki, H. Zhao, J. Cui, and H. Zha, "A novel laser-based system: Fully online detection of abnormal activity via an unsupervised method," in *ICRA*, 2011. 2