RGBD-CAMERA BASED GET-UP EVENT DETECTION FOR HOSPITAL FALL PREVENTION

Bingbing Ni[1], *Nguyen Chi Dat*[2] *and Pierre Moulin*[3]

1. Advanced Digital Sciences Center, Singapore 138632
2. National University of Singapore, Singapore 119077
3. University of Illinois at Urbana-Champaign, IL 61820-5711

ABSTRACT

In this work, we develop a computer vision based fall prevention system for hospital ward application. To prevent potential falls, once the event of *patient get up from the bed* is automatically detected, nursing staffs are alarmed immediately for assistance. For the detection task, we use a RGBD sensor (Microsoft Kinect). The geometric prior knowledge is exploited by identifying a set of task-specific feature *channels*, e.g., regions of interest. Extensive motion and shape features from both color and depth image sequences are extracted. Features from multiple modalities and channels are fused via a multiple kernel learning framework for training the event detector. Experimental results demonstrate the high accuracy and efficiency achieved by the proposed system.

Index Terms— multi-modal, depth image, data fusion, multiple kernel learning, event detection

1. INTRODUCTION

Falls account for up to 70% of accidents among hospitalized patients. The most frequently cited activities at the time of falling is getting up from bedside commodes, transferring from the bed and chair to the bathroom or toilet [1]. Falls cause injury and death for the patients, and risk of falls increases markedly with age. Accelerometers [2] and gyroscopes [3] are used for detecting falls. As a non-invasive technique, computer vision systems are developed to detect accidental falls in elderly home care applications [4]. To actively prevent falls, fall risk assessment, patient-specific prevention plan, educational handout and poster for over the patient's hospital bed, have been recently introduced to hospitals, which has shown to reduce the number of elderly patients with falls in hospitals, according to [5].

The implementation of these fall prevention systems, however, is complex and indirect. One direct and economic solution could be an automatic system which can detect the *pre-fall* event such as *patient gets up from the bed*. Then an alarm can be raised and the nursing staff can come immediately for assistance. In this work, we are interested in employing computer vision techniques for reliably detecting the event *patient gets up from the bed*. Being able to detect and recognize human activity and event is important for performing assistive tasks. Chan et al. use sensor networks to monitor the user in a home environment [6]. However, these sensor based approaches are often cost prohibitive and they may require subjects to wear RFID sensors in an environment labeled with RFID tags, as shown in [7], which is invasive. Non-invasive methods such as computer vision technique have been extensively investigated. One common approach is to use space-time features to model points of interest in video [8, 9]. More recently, dense trajectories are utilized for activity recognition [10]. However, due to the large variations existing in illumination, people's posture, clothing, etc., $2D$ video based methods generally give un-robust and inaccurate detection or recognition results.

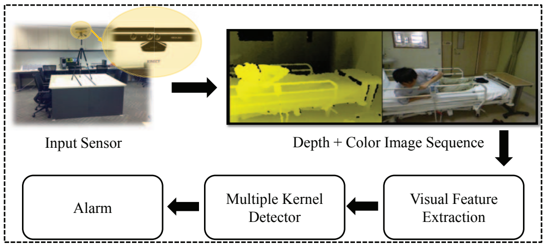

Fig. 1. Overview of the proposed hospital fall prevention system.

Recent emergence of depth sensor (e.g., Microsoft Kinect) has made it feasible and economically sound to capture in real-time not only color images, but also depth maps with appropriate resolution (e.g., 640×480 in pixel) and accuracy (e.g., $< 1cm$). A depth sensor together with a color camera can provide three-dimensional structure information of the scene as well as the three-dimensional motion information of the subjects/objects in the scene, which has shown to be advantageous for action recognitions [11].

In this work, we perform event detection (*patient gets up from the bed*) using an inexpensive RGBD sensor (Microsoft Kinect). Input to the detection system is the synchronized color and depth video streams. Prior domain knowledge of

This work is performed when Nguyen is a research intern in ADSC.

the task is utilized by identifying a set of task-specific feature *channels*, e.g., regions of interest, and multiple motion and shape features are extracted from both color and depth modalities. To fuse features extracted from different channels and modalities, we employ the multiple kernel learning framework [12]. We show that we can achieve highly accurate and robust detection performance and the system operates in real-time. A system overview is illustrated in Figure 1.

2. APPROACH

In this section, we will first introduce our multiple modality visual feature extraction method, by exploring prior domain knowledge of the task. Then, a multiple kernel learning based multiple modality feature fusion method is described for event detection.

2.1. Multi-modal Multi-channel Feature Extraction

Key to visual detection performance is the design of visual features for event representation. Generic feature representations such as bag-of-visual words [13] could be too *coarse* to capture discriminative information. On the other hand, model based methods such as template matching [14] could be too specific to be over-fitting. We note that there exists rich prior domain knowledge for our specific task, e.g., *get up* detection. As the event *get up* mostly occurs around the bed regions, features extracted from these regions are critical to identify such an event. Successful utilization of this important prior knowledge can not only reduce the feature dimensionality but also lead to more discriminative representation. In a well-controlled hospital ward environment, we fix the camera configuration which mounts to a bed from the side view. We therefore manually define a rectangular region of interest enclosing the bed area. To facilitate subsequent feature extraction, we further divided the entire region of interest into $4 \times 2 = 8$ equally-sized rectangular blocks, horizontally and vertically. Each of these blocks, denoted as a *channel* in the rest of this paper, corresponds to bedside, middle, or end of the bed etc., respectively. Note that this method is view angle dependent, in future work, we will explore the geometrical features of the ward (e.g., lines of the bed, wall) to rectify the scene image in order to achieve view angle invariance.

To obtain discriminative features for action detection, we investigate multiple motion and shape features including motion history images (MHI), histogram of oriented gradients (HOG) and histogram of optic flows (HOF). These features characterize human actions from different aspects, which are complementary to each other. In our system, we compute the MHI images throughout frames for each predefined block (channel) and the resulting MHI is down-sampled to the size of 10×10 pixels, i.e., a feature vector of length 100. MHI images are calculated for both color images and depth images based on the algorithm proposed by Bobick and Davis [14]. The updating formula to calculate MHI for the depth channel is given by:

$$H_\tau^D(x,y,t) = \begin{cases} \tau, if(|D(x,y,t) - D(x,y,t-1)|) > \delta D_{th} \\ \max(0, H_\tau^D(x,y,t-1) - 1), else. \end{cases} \quad (1)$$

Here, H_τ^D denotes the motion history image and $D(x,y,t)$ denotes the depth sequence. δD_{th} is the threshold value for generating the mask for the region of motion in the depth direction. To obtain HOG features, each channel is then divided into $2 \times 2 = 4$ cells, and then eight bins of gradient directions are used. For HOF features, nine bins which consist of eight bins of gradient directions and one zero bin are used. We use the same methods as in [10] for calculating HOG and HOF. To form representations for HOG and HOF features over frames (variable length), we use the bag-of-features approach. Note that previous work has also found that the HOF and HOG features may perform very well on human action recognition task [10, 15]. As a summary, given a video clip, we extract six different features, i.e., MHI, HOG, HOF for both RGB and depth sequences, for each of the eight channels. Figure 2 illustrates the feature extraction process. Note that we haven't used multi-scale feature representations (pyramid), since the current single scale feature representation already achieves high performance with great efficiency.

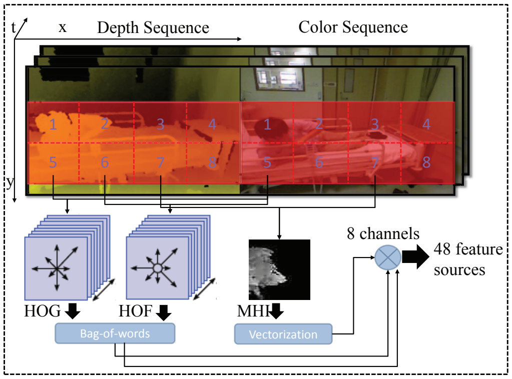

Fig. 2. The feature extraction process. The region of interest is divided into 8 channels. For each channel, MHI, HOG, HOF features are extracted for both RGB and depth modalities. For HOG and HOF features, we use the bag-of-words representations.

2.2. Multiple Feature Fusion for Detection

As seen from the previous subsection, there are 48 feature sources (different type of features, sensor modalities and channels). To optimally combine all these features, in this work, we utilize multiple kernel learning (MKL) framework [12]. The objective in multiple kernel learning (MKL) is to jointly learn both kernel and support vector machine (SVM) parameters, which are regularized to encourage sparse kernel combination. Specifically, the objective is to construct a multiple kernel based classifier $f(\mathbf{x})$ as:

$$f(\mathbf{x}) = \sum_{i=1}^{L} \alpha^i K(\mathbf{x}, \mathbf{x}_i) + b, \quad (2)$$

$$K(\mathbf{x}', \mathbf{x}) = \sum_{m=1}^{M} d_m K_m(\mathbf{x}', \mathbf{x}), \quad (3)$$

$$s.t. \quad d_m \geq 0, \sum_{m=1}^{M} d_m = 1, \quad (4)$$

where, $\mathbf{x}$ denotes a feature vector, K_m denotes $m - th$ kernel. We assume there are L labeled training data. In our work, we endow each HOG and HOF feature with a χ^2 kernel, i.e., $K(\mathbf{x}, \mathbf{x}') = \exp(-\sum_{i=1}^{d} \frac{(x_i - x_i')^2}{x_i + x_i'} / \sigma^2)$, where $\mathbf{x} = [x_1, \cdots, x_d]^T$ and $\mathbf{x}' = [x_1', \cdots, x_d']^T$ are two d-dimensional feature vectors. We also endow each MHI feature with a Gaussian kernel. Note the bandwidth parameter σ^2s are set at the mean of the (squared) distances (χ^2 distance for HOG, HOF and squared Euclidean distance for MHI) of all training feature pairs. The optimization problem could be formulated as:

$$\min_f \quad \frac{1}{2} \|f\|_{\mathcal{H}}^2 + C \sum_i \xi_i, \quad (5)$$

$$s.t. \quad y_i(f(\mathbf{x}_i) + b) = 1 - \xi_i, \xi_i > 0, \forall i, \quad (6)$$

where $\|.\|_{\mathcal{H}}$ denotes the norm in Reproducing Kernel Hilbert Space $\mathcal{H}$. y_i denotes the i-th sample label, e.g., $+1$ or -1. We choose a fixed value of the parameter C (C = 100) empirically. To solve Eqn. (5), we use the *SimpleMKL* algorithm [16], which is proved to be efficient and it converges rapidly compared to other MKL optimization algorithms.

3. EXPERIMENTS

3.1. Dataset Construction

We utilize Microsoft Kinect sensor to construct the *get up* event detection video database. Videos are collected in a hospital single bed ward environment. The sensor is setup approximately 3 meters to the bed mounting at the side view. The resolutions of both color image and depth map are 640 × 480 in pixel. The color image is of 24-bit RGB values; and each depth pixel is an 16-bit integer. Both sequences are synchronized and the frame rates are 30 frames per second (FPS). Camera configuration is fixed throughout the capture session. We capture about $50,000$ frames (approximately 0.5 hours long). We manually crop out the segment of the event *patient gets up from the bed* (positive sample) from the whole video corpus. Each positive sample spans about $5 - 10$ seconds. Negative samples are randomly cropped from the rest of the video, each of which also spans about $5 - 10$ seconds. Finally, we obtain 240 video samples, which consist of 40 positive samples and 200 negative samples, captured from 4 subjects (patients).

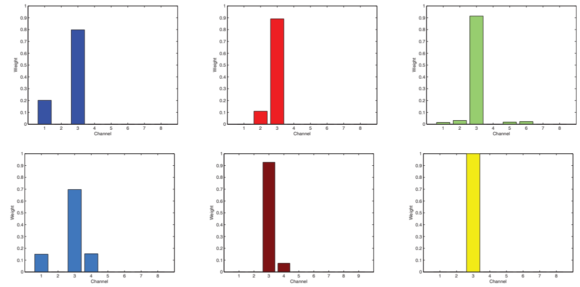

Fig. 3. Contributions (weights) of each type feature from different channels for the trained event detector. Left to right, top to bottom: MHI RGB, HOG RGB, HOF RGB, MHI Depth, HOG Depth, HOF Depth, respectively.

3.2. Experimental Results

For all experiments, we use a leave-one-subject-out testing scheme. Namely, in each run, the video samples from one subject are chosen as testing samples and the rest as training samples.

We first conduct experiments to compare the performances of each type of features, i.e., motion features MHI, HOF and shape features HOG, from both color and depth sensor modality, respectively. We also evaluate the performance when multiple features are combined. The precision/recall curves are compared in Figure 4 and the event recognition accuracies are compared in Table 1.

From Figure 4 and Table 1, we can see that both MHI and HOF features give good performances, but HOG feature is not as capable as these two motion features, which means that motion features are most important for this problem. The combination of different features using MKL gives generally better performances. Even though for this dataset, using only depth MHI features gives slightly better result, we believe that in general, the combination of different features is more robust and can perform much better. We can conclude that MHI features are good enough for this task.

The channel weights d_m when the detector is trained using all feature sources are illustrated in Figure 3. We can notice that the weights are quite sparse which demonstrates the feature selection capability of the MKL framework. Also the most informative features are extracted from channel 3 which are around the end of bed. This well corresponds to the fact that motion occurs in this region is critical for identifying whether the person is getting up. In addition, we can note that HOG and HOF features contribute little when multiple features are combined in MKL.

We also compare our method with the state of the art activity recognition methods including STIPs [8] and dense trajectories [10], as summarized in Table 2 in terms of detection accuracy. Our algorithm significantly outperforms the state-of-the-arts. Note that in the testing, the system operates at the speed of 10 frames per second on a 64-bit Intel i5 core CPU with 6GB memory (using un-optimized C code), which

Table 1. Performances (recognition accuracy) of the event detector by using different features.

Feature	All	MHI RGB	MHI Depth	HOF RGB	HOF Depth	HOG RGB	HOG Depth
Accuracy	98.76	98.35	99.17	95.87	97.52	88.84	86.78

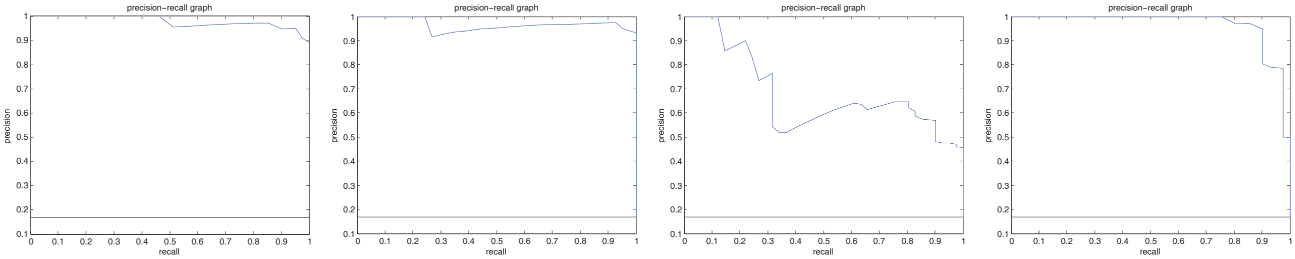

Fig. 4. The precision/recall curves from different features. From left to right: combined features, MHI depth, HOG depth and HOF depth, respectively.

Table 2. Comparisons of performance (recognition accuracy) with the state-of-the-art methods.

Method	STIP	Dense Trajectory	Ours
Accuracy	75.43	85.96	98.76

is real-time.

4. CONCLUSIONS AND FUTURE WORK

We present a vision based fall prevention system based on *get up* event detection for hospital ward monitoring. Combining multiple features from multiple modalities via a MKL framework, the system achieves high accuracy and efficiency. Our future work will focus on constructing more video data and extend the current system to multiple-bed ward, which is more challenging due to scene variations and occlusions.

5. ACKNOWLEDGEMENT

This work is supported by the Human Sixth Sense Project, Illinois at Singapore Pte Ltd.

6. REFERENCES

[1] “http://findarticles.com/p/articles/mim0fsw/is324/ain17213682/.”

[2] A. K. Bourke, C. N. Scanaill, K. M. Culhane, J. V. O’Brien, and G. M. Lyons, “An optimum accelerometer configuration and simple algorithm for accurately detecting falls,” in *IASTED international Conference on Biomedical Engineering*, pp. 156–C160, 2006.

[3] A. K. Bourke and G. M. Lyons, “A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor,” *Medical Engineering and Physics*, vol. 20.

[4] Z. Fu, E. Culurciello, P. Lichtsteiner, and T. Delbruck, “Fall detection using an address-event temporal contrast vision sensor,” in *IEEE International Symposium on Circuits and Systems*, pp. 424–427, 2008.

[5] P. C. Dykes, “Fall prevention in acute care hospitals - a randomized trial,” *The Journal of American Medical Association*.

[6] M. Chan, D. Esteve, C. Escriba, and E. Campo, “A review of smart homescpresent state and future challenges,” *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 55C–81, 2008.

[7] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel, “Inferring activities from interactions with objects,” *Pervasive Computing*, vol. 3, no. 4, pp. 50–C57, 2004.

[8] I. Laptev, “On space-time interest points,” in *IJCV*, vol. 64, pp. 107C–123, 2005.

[9] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *International Workshop on Visual Surveilliance and Performance Evaluation*, 2005.

[10] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, “Action Recognition by Dense Trajectories,” in *IEEE Conference on Computer Vision & Pattern Recognition*, pp. 3169–3176, 2011.

[11] J. Sung, C. Ponce, B. Selman, and A. Saxena, “Human activity detection from rgbd images,” *CoRR*, vol. abs/1107.0169, 2011.

[12] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, “Multiple kernel learning, conic duality, and the smo algorithm,” in *Proceedings of the twenty-first international conference on Machine learning*, ICML ’04, 2004.

[13] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.

[14] A. Bobick and J. Davis, “The representation and recognition of action using temporal templates,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[15] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, “Action dectection in complex scenes with spatial and temporal ambiguities,” in *ICCV*, 2009.

[16] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, “Simplemkl,” *Journal of Machine Learning Research*, 2008.