A Bag of Features Approach to Ambient Fall Detection for Domestic Elder-care

Emmanouil Syngelakis
Centre for Vision Speech and Signal Processing
University of Surrey
Guildford, United Kingdom.
Email: es00034@surrey.ac.uk

John Collomosse
Centre for Vision Speech and Signal Processing
University of Surrey
Guildford, United Kingdom.
Email: jc0028@surrey.ac.uk

Abstract—**Falls in the home are a major source of injury for the elderly. The affordability of commodity video cameras is prompting the development of ambient intelligent environments to monitor the occurence of falls in the home. This paper describes an automated fall detection system, capable of tracking movement and detecting falls in real-time. In particular we explore the application of the Bag of Features paradigm, frequently applied to general activity recognition in Computer Vision, to the domestic fall detection problem. We show that fall detection is feasible using such a framework, evaluted our approach in both controlled test scenarios and domestic scenarios exhibiting uncontrolled fall direction and visually cluttered environments.**

***Keywords*-Ambient domestic monitoring, fall detection, activity recognition, bag of features.**

I. Introduction

Falls are one of the most common and dangerous accidents for older people, especially when living alone. It is estimated that about one third of people aged 65 and above endure a fall at some time each year, and 6 in 10 of these falls occur in the home [1]. These incidents contribute to physical injury and can introduce ongoing psychological problems. The increasing trend toward domestic care, driven by rising costs and an ageing population, motivates new ambient monitoring technologies to detect and alert carers to falls within the home.

This paper reports an experiment into the application of "Bag of Features" (BoF) framework, commonly used in Computer Vision for general activity recognition [2], [3], to the detection of falls in the home. BoF frameworks regularly top the performance tables of international benchmarking competitions in video activity recognition (e.g. TRECVid, VideoOlympics). Such benchmark footage often exhibits visual clutter and moving cameras, yet BoF based approaches are able to identify complex actions such as drinking, kissing, running, etc. in movies. In this paper we apply the BoF paradigm to the problem of domestic fall detection, for which good performance in unconstrained and cluttered domestic environments is essential.

Our passive visual monitoring system uses trained descriptors based on motion patterns rather than shape, and so can, for example, discriminate between video of people lying down and falling. Further our system is robust to the direction of the fall relative to the camera, which can be problematic for ambient monitoring systems relying upon on the outline of the body shape.

We evaluate our system over two datasets each comprising 100 video clips capturing, equally, examples of falls and other normal domestic activities undertaken by 3 individuals. We show mean average precision (MAP) of 90% for uncluttered outdoor visual environments, and similar levels of performance (89%) for cluttered indoor domestic visual environments.

II. Related Work

The most common fall alert systems perform no detection and are based on manually operated panic buttons, worn on the wrist or around the neck. A number of passive monitoring technologies have also been developed, the majority of which are mechanically based sensors. An accelerometer based solution in the form of wrist watch [4] was developed by CSEM (Centre Suisse d'Electronique et de Microtechniuqe). A similar sensing platform (the eWatch [5]) is composed of multiple sensors such as temperature, light, dual axis accelerometers and a microphone. A sensor fusion technique is used to classify user activity in real-time. Orientation of body is also considered in [6], [7]; these systems also consider the possibility of non-horiztonal end positions, for example detecting falls against walls or furniture. These systems underline the importance of using motion information rather than end-body shape of position in determining if a fall has occurred. Typically mechanical sensors yield best results when worn close to the centre of gravity [5]. An alternative form of mechanical sensor is built into a cane in [8], taking measurements of motion, force and pressure. Three stages of fall must be identified to trigger an alarm; a swift change from vertical to horizontal, the detection of force on ground impact, and a motionless cane. These stages are designed to avoid false positive detections when, for example, the cane is dropped.

The majority of vision based fall detection systems rely upon the extraction of a region (silhouette) representing the monitored subject. Analysis of this region's shape indicates a fall. One example of such a system was presented by Lee and Mihailidis [9], using a ceiling mounted camera, which

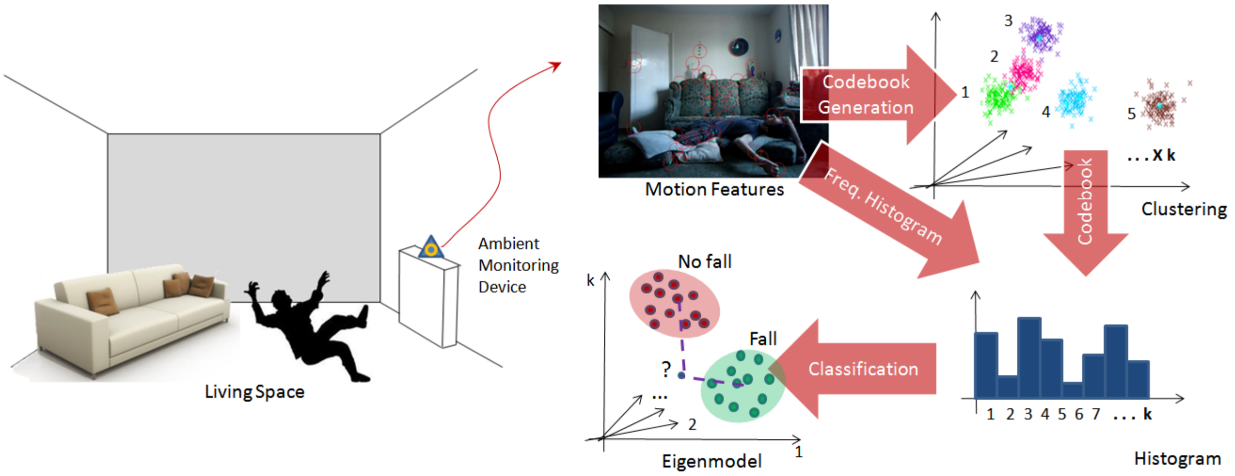

Figure 1. Illustrating the Bag of Features pipeline used in our real-time fall detection system. During training, features are extracted from temporally local windows in the video and clustered via k-means to form a codebook. At run-time the codebook is used to bag features from the current time instant into codewords. The frequency histogram of codewords is classified to determine if the current temporal window contains a fall.

could discriminate between five different action types. Wide-angle [10] lens cameras and omni-directional cameras [11] have also been explored for ambient monitoring. Spehr *et al.* [10] using a hybrid background subtraction techniques to determine the person region invariant to clothes, lighting etc. and used body orientation to determine fall occurance. Specifically a change in orientation of over 28 degrees indicated a fall. A similar orientation based system was presented by Zhang *et al.* [12]. Such shape analysis demands accurate segmentation of the person, and could generate false alarms if a person lies down in the monitored area.

Three dimensional (3D) tracking solutions within the living space have also been explored for fall detection, using calibrated video cameras [13], [14] and time of flight cameras [15]. Rougier and Meunier [13] track the head of a subject using a particle filter and identify falls from the trajectory of the head. The main disadvantage of their lab based prototype was the necessity to manually bootstrap the system by indicating the location of the head. Jansen *et al.* [14] used 3D tracking to distill a set of features comprising head distance to floor, orientation of body and level of activity (motion) to determine occurance of a fall. Although 3D camera systems convey an accurate representation of the environment to the detection algorithm, they require multiple calibrated cameras to be installed within the living space — whereas our system consists of a single camera that can be relocated as desired e.g. on a table or mantlepiece within the living room.

III. Fall Detection System

We now describe the process for detecting falls in real-time. In contrast to typical fall detection systems that isolate the region of interest corresponding to the person, our system requires no such pre-segmentation of the scene. Rather, we detect stable points of interest anywhere in the scene and extract features from the motion of these points recognising the characteristic motion of a fall. By removing the need for identifying the the person in the scene, our system is able to operate in cluttered visual environments that may frustrate segmentation algorithms based on appearance or motion subtraction.

A. *Codebook Extraction*

Our system captures VGA (640×480) resolution video frames, and identifies points of interest within each video frame I_t at time t using the Kanade-Lucas tracker (KLT) across multiple (octave internal) spatial scales [16]. These points are tracked to the subseqeuent frame yielding a set n motion vectors $\mathcal{F}(t) = \{\mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_n}\}$ where $\mathbf{f_i} = \left(\frac{\delta \mathbf{I_t}}{\delta \mathbf{x}}, \frac{\delta \mathbf{I_t}}{\delta \mathbf{y}}\right)$; typically $n = [20, 50]$ resulting in a variable number of features per frame. These features form the basis for detection of falls in our system, being trained to learn and later used recognise the pattern of motion vectors present during a fall.

During training we capture several videos of simulated falls and other unrelated domestic actions, resulting in a large set of motion vectors from each frame of each video. Following the standard BoF hard-assignment method, these features are bagged into "codewords" using unsupervised k-means clustering to form $k = 100$ groups. The cluster centres in the feature space $\mathcal{C} = \{\mathbf{c_1}, \mathbf{c_2}, ..., \mathbf{c_k}\}$ are used to assign each feature in the training data to a codeword $k = \{1..k\}$. We compute a frequency histogram $\mathbf{H(t)} = \{\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_k}\}$ for each training video frame by counting the occurance of codewords present within that frame. Due to the varying number of features in each frame, $\mathbf{H(t)}$ is

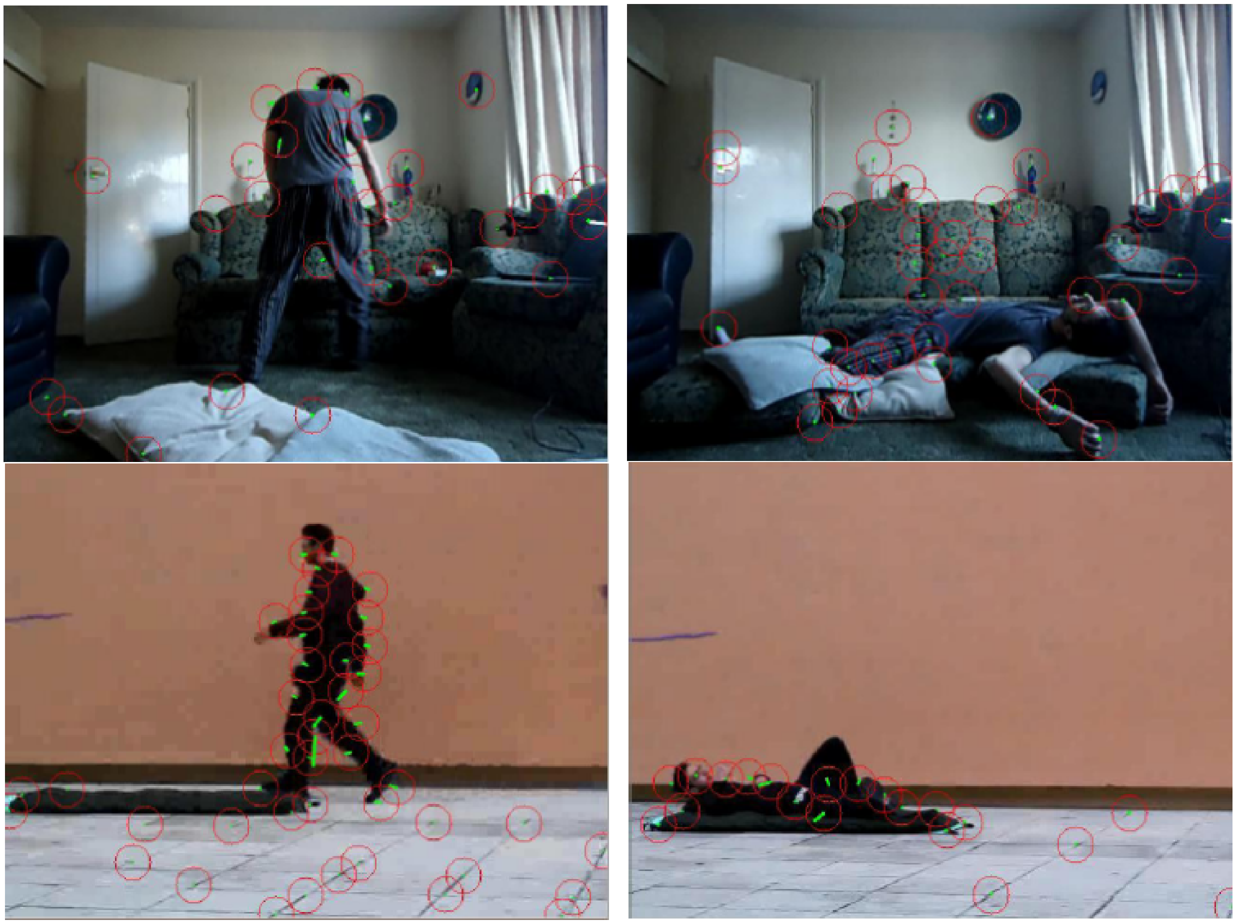

Figure 2. Evaluation dataset: examples of falls present with our two 100 video datasets captured in cluttered and uncluttered conditions. Red circles indicate features; green lines indicate the feature trajectories.

normalised to enable the comparison of histograms between frames.

B. Fall Classification

The training video clips are separated into positive (falls present) and negative (falls absent) categories, and a set of histograms $\mathcal{H}_+$ and $\mathcal{H}_-$ computed using codebook $\mathcal{C}$. Collectively these histograms are points $\{\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_m}\} \in \Re^{\mathbf{k}}$ from which we a mean (μ) and covariance ($\mathbf{C}$).

$$\mu = \sum_{i=1}^{m} \mathbf{p_i} \tag{1}$$

$$\mathbf{C} = (\mathbf{p} - \mu)(\mathbf{p} - \mu)^{\mathbf{T}} \tag{2}$$

from which we compute the eigenvectors $\{\mathbf{u_1}, \mathbf{u_2}, ..., \mathbf{u_k}\}$ and eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_k\}$ of $\mathbf{C}$:

$$\begin{aligned} \mathbf{C} &= \mathbf{UVU^T} \\ \mathbf{U} &= \left[\begin{array}{cccc} \mathbf{u_1} & \mathbf{u_2} & ... & \mathbf{u_k} \end{array} \right] \\ \mathbf{V} &= \left[\begin{array}{ccc} \lambda_1 & & 0 \\ & ... & \\ 0 & & \lambda_k \end{array} \right] \end{aligned} \tag{3}$$

We compute d such that $\sum_{i=1}^{d} \lambda_i \geq 0.95$ creating a projection space $\mathbf{P} = \left[\begin{array}{cccc} \mathbf{u_1} & \mathbf{u_2} & ... & \mathbf{u_d} \end{array} \right]$ from which we compute a reduced dimension representation $\mathcal{H}$ retaining 95% of the variance in the training set:

$$\begin{aligned} \hat{\mathcal{H}}_+ &= \mathbf{P^T}\mathcal{H}_+ \\ \hat{\mathcal{H}}_- &= \mathbf{P^T}\mathcal{H}_- \end{aligned} \tag{4}$$

From the $d-$dimensional training examples $\{\hat{\mathcal{H}}_+, \hat{\mathcal{H}}_-\}$ the respective means $\{\mu_+, \mu_-\}$ and covariances $\{\mathbf{C}_+, \mathbf{C}_-\}$ are computed, so comprising Eigenmodels for the positive and negative fall detection cases.

At run-time, each video frame is classified in real-time as being a positive or negative fall detection case by comparison with these Eigenmodels. First, motion features are extracted and codebook applied to create a query histogram $\mathbf{q} \in \Re^{\mathbf{k}}$, which is subsequently projected into the $d-$dimensional classification space:

$$\hat{\mathbf{q}} = \mathbf{P^T q} \tag{5}$$

The Mahalanobis distance to each Eigenmodel is computed to determine whether a fall has taken place. A fall has occured if:

$$(\mathbf{q}-\mathbf{u}_{+})\mathbf{C}_{+}^{-1}(\mathbf{q}-\mathbf{u}_{+})^{\mathbf{T}} < (\mathbf{q}-\mathbf{u}_{-})\mathbf{C}_{-}^{-1}(\mathbf{q}-\mathbf{u}_{-})^{\mathbf{T}} \qquad (6)$$

We low-pass filter this per-frame decision on fall presence by counting the occurance of falls within a short temporal window (10 frames). If more than half of these frames are deemed to contain falls then the alarm is triggered.

IV. Experimental Results

We evaluted our system over two datasets each containing 100 video clips split evenly between fall and non-fall scenarios. The datasets were created using three participants. Non-fall scenarios included a random spread of activities likely to occur in domestic scenarios such as sitting and reclining on couches and ambualtion within the visual field of the camera. In the case of the cluttered dataset, footage was captured in a living room with typical visual clutter. To enable participants to realistically simulate falling, cushions were placed on the floor of the capture area. In the uncluttered dataset, fall direction was lateral with respect to the camera. In the cluttered dataset, fall direction was unconstrained and occured in all directions; left, right, towards and away from the camera.

The performance of the system is summarised in the confusion matrix of Table I. For the cluttered and uncluttered scenarios the mean average precision (MAP) was 89% and 90% respectively, with a split of 20:80 between training and classification test data. The 20% of the dataset used for training was selected at random.

The optimized detection of KLT tracker using the OpenCV library, and the relatively few multiplications and additions requried to compute (6) enables processing of video in real-time at a sustained rate of 20 frames per second on a Pentium 4 2.2Ghz laptop with 2Gb RAM. In our experiments video was captured using the low cost embedded web camera of the laptop.

V. Discussion and Conclusion

We have demonstrated that a Bag of Features (BoF) framework can be effectively applied to the problem of fall detection in real-time, even in the presence of visual clutter and unconstrained fall direction. In contrast to visual detectors based on segmenting and analysing the contour, our system is based on the learned motion signatures of falls. This enables us to be robust to visual clutter and distinguish between individuals lying down voluntarily in a controlled manner or falling quickly.

We were surprised at the sustained strong performance of the system when switching from the uncluttered to the cluttered dataset, despite the simiplicity of our features (comprising motion flow only). Future work will explore the possibility of including appearance information in the feature vector, as this has recently been shown [3] to enhance the performance of more general BoF based activity recognition.

	Results	
	+ve	-ve
G.Truth		
+ve	0.92	0.08
-ve	0.16	0.84

	Results	
	+ve	-ve
G.Truth		
+ve	0.84	0.16
-ve	0.6	0.94

Table I

Precision of the system on the cluttered (left) and uncluttered (right) datasets (80 videos per dataset).

A further refinement could also be in the deployment of our system. Currently we employ a laptop with integrated web camera, running our software in real-time. In future development we could leverage small form factor devices specifically built to co-locate within the shared living space — much as modern Carbon Monoxide and smoke alarms do. However we do not believe such enhancements are necessary to further demonstrate the potential of deploying BoF for real-time fall detection in low-cost ambient monitoring solutions.

References

[1] M. Sartini, M. Cristina, A. Spagnolo, P. Cremonisi, C. Costaguta, F. Monacelli, J. Garau, and P. Odetti, "The epidemiology of domestic injurious falls in a community dwelling elderly population: an outgrowing economic burden," *Europeon Journal of Public Health*, 2009.

[2] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proc. Intl. Conf on Pattern Recognition (ICPR)*, vol. 3, 2004, pp. 32–36.

[3] K. Mikolajczyk and H. Uemura, "Action recognition with appearance motion features and fast search trees," *Computer Vision and Image Understanding (CVIU*, 2011.

[4] "Fall detection system (centre suisse d'electronique et de microtechnique)," 2010, http://www.csem.ch/docs/Show.aspx?id=6026, Last accessed November 2010.

[5] J. Chen, K. Kwong, D. Chang, J. Luk, and R. Bajesy, "Wearable sensors for reliable fall detection," in *Proc. IEEE Engineering in Medicine and Biology*, 2005.

[6] U. Maurer, A. Smailagic, D. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body," in *Proc. Intl. Workshop on Wearable and Implantable Body Sensor Network Positions (BodyNet)*, 2006.

[7] "Lifeline with autoalert (phillips electronics)," 2010, http://www.newscenter.phillips.com/main/standard/news/press/2010/20100322_lifeline.wpd, Last accessed November 2010.

[8] M. Lan, A. Nahapetian, A. Vahdatpour, L. Au, and M. S. W. Kaiser, "Smartfall: An automatic fall detection system based on subsequence matching for the smartcane," in *Proc. Intl. Workshop on Wearable and Implantable Body Sensor Network Positions (BodyNet)*, Apr. 2009.

[9] T. Lee and A. Mihailidis, “An intelligent emergency response system: preliminary development and testing of automated fall detection,” *Journal of Telemedicine and Telecare*, vol. 11, pp. 194–198, 2005.

[10] J. Spehr, M. Govercin, S. Winkelbach, E. Steinhagen-Thiessen, and F. Wahl, “Visual fall detection in home environments,” in *Proc. Intl. Conf. of the Intl. Society of Gerontology*, vol. 7, no. 2, Jun. 2008, p. 114.

[11] Y. C. Huang, S. G. Miaou, and T. Y. Liao, “A human fall detection system using an omni-directional camera in practical environments for health care applications,” in *Proc. IAPR Conf. on Machine Vision Applications*, 2009.

[12] Z. Zhang, E. Becker, R. Arora, and V. Athitsos, “Experiments with computer vision methods for fall detection,” in *Proc. Intl. Conf. on Pervasive Tech. related to Assistive Environments*, 2010.

[13] C. Rougier and J. Meunier, “Fall detection using 3d head trajectory extracted from a single camera video sequence,” in *Proc. Intl. Workshop on Video Processing for Security (V4PS-06)*, Quebec, Canada, 2006.

[14] B. Jansen and R. Deklerck, “Home monitoring of elderly people with 3d camera technology,” in *Proc. BENELUX Biomedical Engineering Symp.*, 2006.

[15] G. Diraco, A. Leone, and P. Siciliano, “An active vision system for fall detection and posture recognition in elderly healthcare,” in *Proc. Conf. on Design, Automation and Test in Eurpoe*, Mar. 2010, pp. 1536–1541.

[16] B. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proc. Intl. Joint Conf. on Artificial Intelligence (IJCAI)*, 1981, pp. 674–679.