Thermal Imaging Based Elderly Fall Detection

Somasundaram Vadivelu[1], Sudakshin Ganesan[1], O.V. Ramana Murthy[1(✉)], and Abhinav Dhall[2]

[1] Department of Electrical and Electronics Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India
ovr_murthy@cb.amrita.edu

[2] Department of Computer Science, University of Waterloo, Waterloo, Canada

Abstract. Elderly fall detection is very special case of human action recognition from videos and has very practical application in old age home and nursing centers. Fall detection in its simplest form is a binary classification of fall event or other daily routine activities. Hence, the current trend of sophisticated techniques being developed for human action recognition, particularly with scenarios of large number of classes may not be required in elderly fall detection. However, other design considerations such as simplicity (ready to be deployed), privacy issues (not revealing the identity) are to focused and are the major contributions of this paper. The Spatio-Temporal Interest Points (STIP) and Fisher vector framework for human action recognition is established as baseline in this work. A novel optical flow based technique is proposed that yields better performance than the baseline. Further, a very economical thermal imaging based input modality is proposed. Along with the thermal images not revealing the identity of the persons, thermal images also aid human detection from backgrounds – a useful solution in computing the optical flow of human movements. The proposed solution is also validated on the **KUL Simulated Fall** dataset showing its generalization capability.

1 Introduction

Statistically, about one third of the population over the age of 65, tend to fall at least once a year[1], and the risk of falls increases with further ageing. Often, such falls can lead to injury and death in extreme cases among the elderly. It has been identified that 87% of the fractures in the elderly are primarily due to the falls. Fall incidents lead to hospital or nursing home admissions, making it further expensive in charges. Sadly, 40% of such admitted do not return to independent living and worse, one-fourth die within a year. Even if there are no injuries sustained from a fall, a large percentage of fallers (47%) cannot get up without assistance. The gap period of time spent immobile between a fall and assistance received and crucial. Longer the gap the higher is the risk of health

S. Vadivelu and S. Ganesan—Equal contribution

[1] http://www.alohahomecarefl.net/orthopedic-care/fall-reduction-and-home-safety/?.

C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part III, LNCS 10118, pp. 541–553, 2017.
DOI: 10.1007/978-3-319-54526-4_40

affects consequence such as dehydration, hypothermia, pneumonia and pressure sores. Apart from physical injuries, fall can also lead to psychological issues such as fear of future falls, feeling lonely and depression [1]. Getting immediate help after a fall largely reduces many such health risks and also enhances the confidence to return to independent living.

Timely detection of falls by automation can be used to trigger alarm to the caring personnel. Further, the past fall incidents can be analyzed to avoid similar situations leading to falls in the future. Often, the fall incidents also aid in identifying health related issues that happened to be unnoticed by the subject until then. In this way the quality of the elderly can be enhanced.

There are three major approaches for automatic fall detection as follows

1. Wearables. These are body-worn containing sensors such as MEMS accelerometer [2]. The data being collected directly from the boy contact, is rich and yield very high accurate performance. However, these devices need to be always worn all the time. Usually, old people tend to forget things easily and further the wearable requires regularly constant recharge; posing a challenge to the user.
2. Ambient analyzers. These can be pressure sensors installed on the floors [3]. These are non-invasive and also conceal the person's identity. However, it is costly – requiring installation on all the floor. Further, false alarm rate is high due to any nonhuman object falls.
3. Vision/camera based approach [4]. They are non-invasive and economical. However, their solutions can be computationally complex and further the privacy of the individual is not ensured.

In human action recognition (HAR), only one event – fall or one non-fall event – is assumed to constitute the entire segment. However, in the current task, non-fall video segment contains more than one non-fall event such as walking, sitting down, arranging bed and so on; hence challenging than HAR. In this paper the two limitations of the vision based approaches are addressed as follows

– Thermal imaging based input modality to conceal the individual's identity.
– Optical flow based feature extraction.

Thermal imaging captures human data even during the night or absence of very less ambient light; times when probability of fall is high. The cost of Thermal cameras is also decreasing and we have selected the most economical – Flir ONE thermal imaging camera – as to make our solution practically viable. We have also created a Thermal imaging Fall event dataset which will be released to the research community for further progress in fall detection. The layout of the rest of the paper is as follows. Section 2 contains a literature survey on vision based fall detection techniques and the datasets existing in the literature. Section 3 contains the overall framework and our proposed optical based solution. The datasets used to conduct our experiments is discussed in Sect. 4. Section 5 contains the results obtained by applying our proposed techniques and comparisons with baseline framework of human action recognition. Finally we summarize in Sect. 6.

2 Related Literature

This section contains a brief survey on the vision based techniques and the datasets proposed in the literature to conduct simulation studies.

2.1 Vision-Based Techniques

A good survey on fall detection systems was recently given in [5]. Being non-invasive and economical set-up, camera based solutions are attracting vast majority of researchers. Debard *et al.* [4] used background subtraction and a particle filter to get best estimate for the background. The person is then located in the frame by fitting the biggest ellipse in the foreground section of the frame. The particle filter was based on three different measuring functions computed using – an ellipse that fitted the foreground binary image best, histogram correlation of the ellipse divided into four different parts of the body; and an upper body detector. Any portions inside the fitted ellipse were updated very slowly while the rest of the image (as background) was updated very fast. When the person location in the video frame is thus finalized, a fall feature vector is comprised consisting of values (fitted ellipse's aspect ratio, change in aspect ratio, fall angle, centre speed and head speed) calculated over different time slots before, during and after the fall incident. The feature vector is subsequently used for learning a Support Vector Machine (SVM) model to discriminate fall from non-fall scenario. Auvinet *et al.* [6] computed a set of features extracted from human body silhouette tracking. The features are based on height and width of human body bounding box, the users trajectory with her/his orientation, Projection Histograms and moments of order 0, 1 and 2. They investigated several combinations of usual transformations of the features (Fourier Transform, Wavelet transform, first and second derivatives). Charfi *et al.* [7] detected fall events by analyzing the volume distribution along the vertical axis. An alarm is triggered when the major part of this distribution is abnormally near the floor during a predefined period of time, which implies that a person has fallen on the floor.

2.2 Datasets

Due to ethical and privacy related reasons, the videos of real fall incidents are not shared to research public at large. Hence, majority of the algorithms are evaluated on simulated falls performed in artificial surroundings [4]. However, Kangas *et al.* [8] reported significant differences between falls in artificial surroundings and real fall incidents in terms factors such as illumination and occlusions. Auvinet *et al.* [6] shared dataset containing video segments ranging from 30 s to 4 min using eight calibrated cameras. The video segments contain 22 fall events and 24 other events such as sitting, crouching and lying on a sofa. Charfi *et al.* [7], used a single camera setup to record 249 video segments. These video segments ranged from 10 to 45 s. The datasets contained 192 fall video segments and 57 non-fall/normal activities such as sitting down, walking, standing up and

housekeeping. Baldewijns *et al.* [9] created and shared publicly **KUL Simulated Fall** dataset consisting of 55 fall events and 14 normal daily activities. This dataset was created by re-enacting real-life fall incidents that were observed during previous studies [10].

A limitation observed with all these techniques is the generalization capacity of their proposed techniques is not validated i.e. not tested on multiple datasets. To avoid situations of overfitting to one dataset, we investigate state-of-the-art human action recognition framework and propose novel optical flow technique on two datasets with completely different settings – RGB and Thermal videos.

3 Overall Framework and Background

The overall layout of the proposed framework is shown in Fig. 1.

Fig. 1. Overall framework: starting with an input video segment, optical flow vectors are computed w.r.t the first frame (taken as background image). Angles of the flow vectors are computed and categorized into 10 bins. The distribution of each bin over time is analysed in frequency domain to generate a temporal feature vector. Finally, a SVM classifier is used to predict the action label in the given video segment – fall or non-fall.

The action recognition framework of Fisher vector computation from spatio-temporal interest points is taken as baseline to compare performance of our proposed technique. Firstly, interest points, (Harris 3D corners), are detected. Local descriptors are computed around these detected interest points. Gaussian Mixture Modelling is applied to each of these descriptors to yield GMM centres. Fisher vectors are generated from these GMMs to learn a classifier (for each action class). This baseline benchmark is described below.

3.1 Harris 3D Corner Interest Points

Laptev and Lindeberg [11] proposed the Harris 3D corners as an extension of traditional (2D) Harris corner points for spatio-temporal analysis and action recognition. These interest points are local maxima of a function of space-time gradients. They compute a spatio-temporal second-moment matrix at each video point in different spatio-temporal scales. The interest points are obtained as local maxima of a function of this second-moment matrix. We use the original implementation[2] with standard parameter settings. The interest points are extracted at multiple scales based on a regular sampling of spatial and temporal scale

[2] http://www.di.ens.fr/~laptev/download.html/#stip.

values. They are defined in 5 dimensions (x, y, t, σ, τ), where x, y and t are spatial and temporal axes, respectively; while σ and τ are the spatial and temporal scales, respectively. We then compute the local descriptors – histograms of oriented gradients (HOG) and histograms of optical flow (HOF) with default parameter setting. While the former captures the local motion and appearance, the latter captures the temporal changes.

3.2 Feature Encoding

We construct Fisher vectors (FV) [12] for each HOG and HOF descriptors separately. Initially, 250,000 descriptors are randomly from the training set Gaussian Mixture Modeling (GMM) is applied. We define the parameters obtained from the GMM fitting as $\theta = (\pi_j, m_j, \sum_j; j = 1, 2, ..., k)$ where π_j, m_j and $\sum_j$ are the prior probability, mean and covariance of each distribution. The GMM ascribes each descriptor x_i to a mode j in the mixture with a strength given by its corresponding posterior probability

$$q_{ij} = \frac{(x_i - m_j)^T \sum_j^{-1} (x_i - m_j)}{\sum_{t=1}^{k} (x_i - m_t)^T \sum_t^{-1} (x_i - m_t)}. \tag{1}$$

The mean (u_{jk}) and deviation vectors (v_{jk}) in each mode k are computed as follows

$$u_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \frac{x_{ji} - m_{ik}}{\sigma_i} \tag{2}$$

$$v_{jk} = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} q_{ik} [(\frac{x_{ji} - m_{ik}}{\sigma_i})^2 - 1] \tag{3}$$

where $j = 1, 2, ..., D$ spans the local descriptor vector dimensions. The resultant vectors (u_{jk}) and (v_{jk}) for each of the k modes in the Gaussian mixtures are concatenated to yield w, and then normalised by the 'power-law normalisation' [13] defined as $w_j = |w_j|^\alpha \times sign(w_j)$ with $\alpha = 0.5$. Finally, the vector is L_2-normalised as $w = \frac{w}{||w||}$ to yield the FV vector. We use $k = 256$ for each descriptor type in all our experiments.

For classification, we concatenate all Fisher Vectors (of different descriptors) and use linear SVM LIBLINEAR [14] to select the class with the highest score.

3.3 Frequency Analysis of Optical Flow Angle Bins

In this section, we propose a novel approach to model the temporal information of the optical flow angular bins. We take motivation from the recent work on temporal modelling in human action recognition in [15]. We compute optical flow of each frame w.r.t to the background and hypothesize that angles of optical flow vector will hold discriminating information for a fall action from non-fall action. Hence, we compute the angles of the optical flow vectors and assign them into

10 bins; $0-360^0$ is divided into 10 bins each of size 36^o. Thus, 10 histograms are obtained for each frame optical flow vector. The distribution of each histogram intensity over time dimension contains the action information in the given video segment. We then apply frequency analysis to derive the feature vector from that time distribution curve.

We perform frequency analysis in Fourier transform domain and as outlined in Algorithm 1. Initially, the temporal occurrence of each angular bin (of the optical flow vector) over entire duration of video is collected as a distribution-time curve. Fourier transform [16,17] is applied to detect which frequencies occur most. High frequencies are indicative of large angular changes in the optical flow vector indicating actions such as sudden fall of people. Thus, we collect amplitudes of the constituent frequencies and use it to learn a classifier.

Algorithm 1. Computing temporal feature vectors from 10 Angular bins of optical flow

Input: Video
Output: Temporal feature vectors for the video
$N \leftarrow$number of frames in the video
$I_{Bgrnd} \leftarrow$First frame
for $j \leftarrow 2$ ***to*** N **do**
  $Opt =$Optical flow vector w.r.t I_{Bgrnd}
  $Ang =$Angles of the Opt
  $Bin[j][0...1] =$ Divide Ang into 10 bins
for $i \leftarrow 1$ ***to*** *10* **do**
  /* Apply N-point Fast Fourier Transform */
  $Y = \frac{1}{N}||fft(Bin[][i])||^2$
  /* Omit the d.c component; Select the first $\frac{N-1}{2}$ values */
  /* Rearrange amplitude values in descending manner */
  $[a, b] = descendsort(Y)$
  /* Select amplitude components only as the Temporal feature vector */
  $TFA = a$
return *Temporal Feature Vectors* TFA

4 Datasets

The proposed technique is applied on two datasets – benchmark **KUL Simulated Fall** [9] dataset and in-house created **Thermal Simulated Fall** dataset. The details are given below

KUL Simulated Fall dataset: This dataset contains a total of 55 simulated fall scenarios and 17 video segments containing non-fall activities. Real-life fall incidents recorded and described as in [4] were studied and then re-enacted by 10 different actors. A room was set-up similar to a nursing-home room environment.

Table 1. Overview of different fall scenarios in **KUL Simulated Fall dataset**

Using walking aid	Scenarios
Walker	13
Wheelchair	4
No walking aid	38
Fall speed	
Slow falls	22
Fast falls	33
Moving objects during the fall	
Walker	8
Wheelchair	4
Blanket	3
Chair	10
None	30
Starting pose	
Standing	26
Sitting	10
Bending over	5
Squatting	6
Transitions (sit stand)	8
Ending positions	
Lying on the floor	49
Getting back-up after fall	6

Five IP-cameras (12 fps, resolution: 640×480) were installed in this room to record fall and normal daily activities. The average length of fall event is around 2 min 45 s, with a minimum length of 50 s and a maximum length of 4 min 58 s. Fall scenarios differed from each other in terms of using walking aid, other moving objects during the fall, fall speed, starting pose and ending pose. Different types of fall scenarios are summarized in Table 1.

The video segments containing non-fall activities had an average length of 20 min 39 s per scenario of which the shortest segment was 11 min 38 s and the longest was 35 min 30 s duration. Each segment consisted of a set of normal activities. Non-fall segments varied from each other in terms of using different walking aids, changing the order in which the activities were performed, changing the pace of the person performing the actions, not including all activities in each scenarios and finally not performing all activities in the same positions (e.g. sleeping on top of the blankets, under the blankets or in the chair). The normal activities included were as follows [9]:

– Walking (with and without walking aid);
– Transitions from sitting to standing and vice-versa;

– Sitting;
– Eating and drinking;
– Getting into and out of bed;
– Sleeping;
– Changing clothes;
– Removing and putting on shoes;
– Reading;
– Transfers from wheelchair to chair and vice-versa;
– Making the bed;
– Coughing and sneezing violently;
– Picking something up from the floor.

Thermal Simulated Fall dataset. This dataset is based on **KUL Simulated Fall** dataset. It contains 9 video segments with non-fall scenarios and 35 segments with fall scenarios. These videos are having spatial resolution of 640×480 and captured in a room setting, single view using FLIR ONE thermal camera mounted to a Android phone[3]. FLIR ONE is very economical and explored towards practical deployment. Thermal camera has the advantages of protecting the privacy/identity of the individual, capturing even during no-light conditions. Sample snapshots of Real-incident, simulated (RGB) [9] and contributed Thermal dataset are show in Fig. 2. It can be easily observed that the human detection is easier in thermal imaging. This dataset is made publicly available[4].

5 Results and Discussions

As per the experimental settings set forth in [4,9], 10-fold cross-validation is used in each camera, each technique to compute the performance metrics – the precision-recall curves. To draw these performance curves, we need two parameters: (1) Precision and (2) Recall. In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance (Table 2).

The precision curve is obtained by varying the margin value (the distance between the nearest point and the hyperplane in SVM). As the margin value is varied from 0 to 1 in very small increments of 0.01, the values of TP,TN,FP and FN are computed for each margin value. From these values, Precision and Recall are computed for each margin value as follows

[3] http://www.flir.com/flirone/android/.
[4] https://drive.google.com/open?id=0ByBHFkIRDnx6S2M2WllKaVg5eGc.

Fig. 2. Beginning of fall (a) Real-life (b) Enacted RGB, (c) Enacted, thermal; Middle of fall (d) Real-life (e) Enacted, RGB (f) Enacted thermal. Source of (a), (b), (d), (e) [9]

Table 2. AUC computation

Predicted actual	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

The results obtained are shown in Fig. 4 and Table 3.

Table 3. Performance (AUC) of the proposed approach

Approach	Cam1	Cam2	Cam3	Cam4	Cam5	Thermal
Proposed approach	0.71	0.69	0.67	0.65	0.67	0.64
STIP + FV	0.71	0.62	0.65	0.64	0.61	0.60
Baldewijins *et al.* [9]	0.56	0.35	0.40	0.56	0.38	-

5.1 On the Performance of Different Techniques

Baldewijins *et al.* [9] performed the least in all the 5 camera views. Their technique is based on ellipse fitting to human detected in each frame. During occlusions or partial view, human detection fails completely and hence their least performance. The state-of-the-art framework used for human action recognition –STIP+FV – performed better than their technique by 0.08–0.27% (absolute). This technique is based on space-time interest points and not dependent on human localization. Thus they are free from occlusions and partial views; hence their better performance than Baldewijins *et al.* [9]. Our proposed technique is based on optical flow w.r.t initial frame as background image. This technique is based on global motion changes and hence contains more information that the STIP+FV framework, which is dependent on local information (STIP). Hence better than STIP+FV performance by 0–0.07% (absolute). Consistent better performance is observed on **Thermal Simulated Fall** dataset too, supporting the generalization capability of the proposed technique.

Further there is no need of any codebook construction unlike STIP+FV. When any new daily activity has to be considered, STIP+FV framework needs to rebuild codebook; our technique doesn't need such.

5.2 On the Choice of 'N' While Computing FFT

The number of points for computing N-FFT is taken as 1500 because the minimum length of all the videos in the dataset was around 50 s; with frame rate

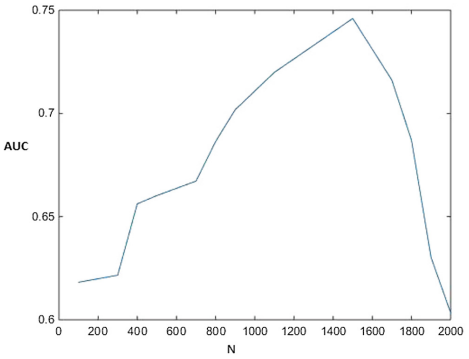

Fig. 3. AUC for different values of 'N' while computing N-point FFT

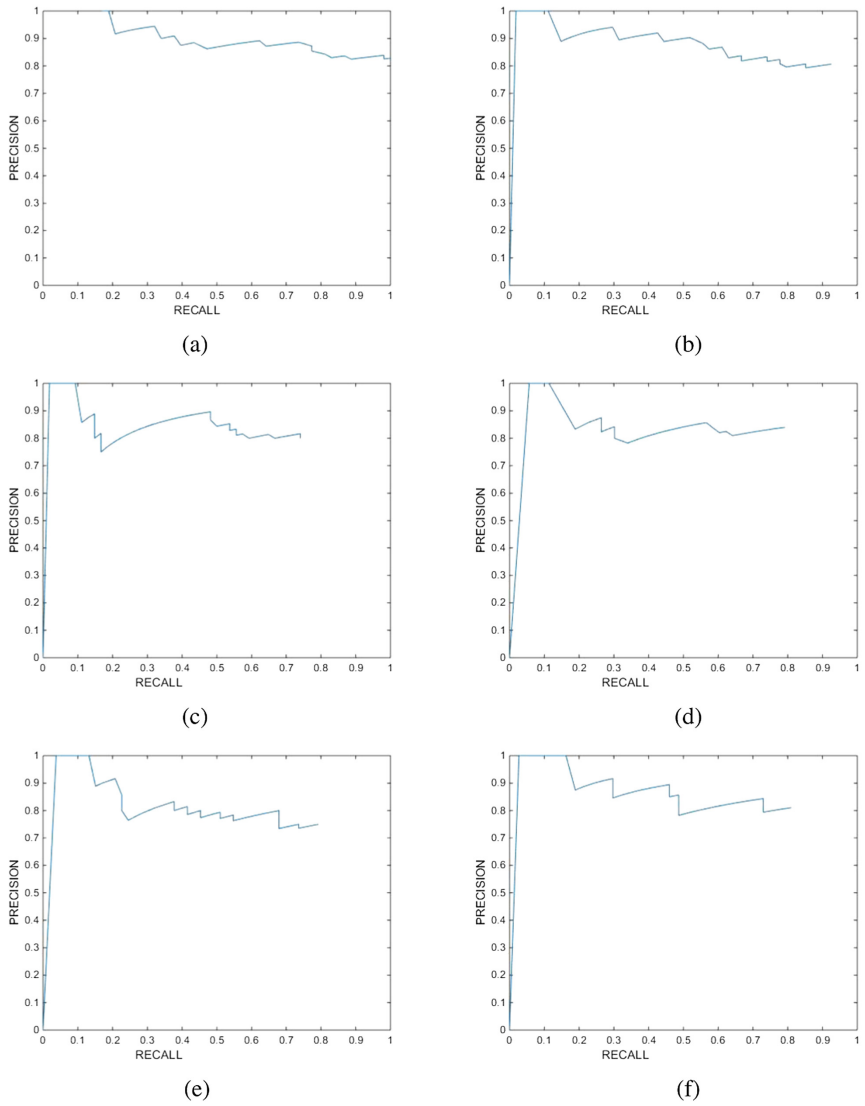

Fig. 4. AUC for (a) Cam 1, STIP (b) Cam 1, proposed technique, (c) Cam 2, proposed, (d) Cam 3, proposed, (e) Cam 4, proposed and (f) Thermal camera, proposed technique.

as 30 fps; totaling to 1500 frames. To investigate further, we also repeated the experiment with our proposed technique on Cam 1 for different values of N. The AUC obtained in each case is plotted as shown in Fig. 3. It can be observed that best AUC is obtained when 'N' is chosen properly according to the video segment length and frame rate of the dataset. In other words, 'N' is dependent on the length of the video segment/step for which we want to compute the decision continuously in real-time implementation.

6 Conclusions

In this paper, we investigated the elderly fall detection using vision based techniques. Skipping the codebook and tracking based techniques, we proposed a simple optical flow based technique. In this technique, angles of the optical flow vectors are computed and categorized into 10 angular bins. The distribution of each bin over temporal dimension is analyzed in frequency domain and a temporal feature vector is constructed for (SVM) classifier training. To conceal the identity of the individual, the proposed solution is validated on a created thermal imaging dataset. The generalization capacity of the proposed technique is thus tested on two datasets. In future, we would like to investigate optical flow w.r.t previous frames and include more daily/routine activities to build enhanced fall detection model.

References

1. Fleming, J., Braynel, C.: Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. Br. Med. J. (BMJ) **337**, 1279–1282 (2008)
2. Bagal, F., Becker, C., Cappello, A., Chiari, L., Aminian, K., Hausdorff, J.M., Zijlstra, W., Klenk, J.: Evaluation of accelerometer-based fall detection algorithms on real-world falls. PLoS ONE **7**, 1–9 (2012)
3. Rimminen, H., Lindstrm, J., Linnavuo, M., Sepponen, R.: Detection of falls among the elderly by a floor sensor using the electric near field. IEEE Trans. Inf. Technol. Biomed. **14**, 1475–1476 (2010)
4. Debard, G., et al.: Camera-based fall detection on real world data. In: Dellaert, F., Frahm, J.-M., Pollefeys, M., Leal-Taixé, L., Rosenhahn, B. (eds.) Outdoor and Large-Scale Real-World Scene Analysis. LNCS, vol. 7474, pp. 356–375. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34091-8_16
5. Mubashir, M., Shao, L., Seed, L.: A survey on fall detection: principles and approaches. Neurocomputing **100**, 144–152 (2013)
6. Auvinet, E., Multon, F., Saint-Arnaud, A., Rousseau, J., Meunier, J.: Fall detection with multiple cameras: an occlusion-resistant method based on 3-D silhouette vertical distribution. IEEE Trans. Inf Technol. Biomed. **15**, 290–300 (2011)
7. Charfi, I., Miteran, J., Dubois, J., Atri, M., Tourki, R.: Definition and performance evaluation of a robust SVM based fall detection solution. In: 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems (SITIS), pp. 218–224 (2012)
8. Kangas, M., Vikman, I., Nyberg, L., Korpelainen, R., Lindblom, J., Jms, T.: Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects. Gait & Posture **35**, 500–505 (2012)
9. Baldewijns, G., Debard, G., Mertes, G., Vanrumste, B., Croonenborghs, T.: Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms. Healthc. Technol. Lett. **3**(5), 6–11 (2016)
10. Vlaeyen, E., Deschodt, M., Debard, G., Dejaeger, E., Boonen, S., Goedemé, T., Vanrumste, B., Milisen, K.: Fall incidents unraveled: a series of 26 video-based real-life fall events in three frail older persons. BMC Geriatr. **13**, 103 (2013)

11. Laptev, I., Lindeberg, T.: Space-time interest points. In: International Conference on Computer Vision (ICCV), pp. 432–439 (2003)
12. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. IEEE (2007)
13. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_11
14. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
15. Murthy, O.V.R., Goecke, R.: The influence of temporal information on human action recognition with large number of classes. In: International Conference on Digital Image Computing: Techniques and Applications (DICTA) (2014)
16. Fourier, J.B.J.: Thorie Analytique de la Chaleur (1822)
17. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex fourier series. Math. Comput. **19**, 297–301 (1965)