A generic interpretable fall detection framework based on low-resolution thermal images

Yannick Wend Kuni Zoetgnande, Jean-Louis Dillenseger

A generic interpretable fall detection framework based on low-resolution thermal images

Yannick Wend Kuni Zoetgnande, Jean-Louis Dillenseger
{yannick.zoetgnande, jean-louis.dillenseger}@univ-rennes1.fr

Univ Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France

Abstract. In this paper, we addressed the particularly challenging problem of fall detection using very low resolution thermal images. We proposed a new method for fall detection only based on the matches and a determined threshold. By classifying a pair of matched points on the ground or not on the ground, we could easily determine how many percent of the shape of a person is on the ground. Thus, we could determine if there is a fall or not. The experiments show that the method is able to classify features of human silhouette as one the ground or not on the ground.

Keywords: Thermal images; Fall detection; Stereo vision; Deep learning

1 Introduction

According to the French National Institute of Demography, the number of elderly people will continue to increase and will double by 2050 (for example, the number of people over 75 years old will increase from 6 million in 2020 to 12 million in 2050). In this context, even if infrastructures are created to welcome them, it becomes essential to find solutions allowing the elderly to live at home, as long as possible and with as much autonomy as possible.

In this context, falls are to be monitored in particular, since they are the first cause of mortality, apart from disease, for individuals over 75 years of age and therefore represent a real societal issue. This surveillance concerns two aspects: detection and prevention of falls. Detection of a fall allows for rapid assistance to be given to the individual. Prevention reduces the number of falls and delays the onset of the first fall. In our case, prevention consists in monitoring the activity of the person and deducing signs of fragility by analyzing the evolution of this activity.

One solution is to propose a new low-cost device based on thermal sensors to prevent the risk of falls by analyzing the activity of seniors. These types of sensors have been chosen because they meet different characteristics of acceptability by the monitored persons : the device is purely passive, these sensors allow a day and night operation and ensure the anonymity of the observed persons.

As the price of the device is an important criterion for the deployment in the living places of the elderly, it was chosen to use low-cost thermal cameras. The main drawback, however, is that low-cost cameras have poor image resolution (80×60 pixels in our case), and the images themselves are

poor in information. The information is purely two-dimensional, yet fall detection and/or activity tracking require an accurate estimation of the person's pose in the room. So we decided to associate two cameras in stereoscopic way.

Our contributions are two-fold:

- A complete interpretable framework that allow to performs 3D vision from very low resolution thermal image
- An hybrid solution (vanilla machine learning + deep learning) for fall detection

The papers is structured ad follows: Section 2 describes the device we used to detect falls and recognise activities, Section 3 presents a new calibration method for thermal cameras, Section 4 details our stereo-vision pipeline with features extraction and matching, Section 5 presents a generic framework for fall detection and Section 6 concludes the paper and presents some perspectives.

2 Stereo vision device

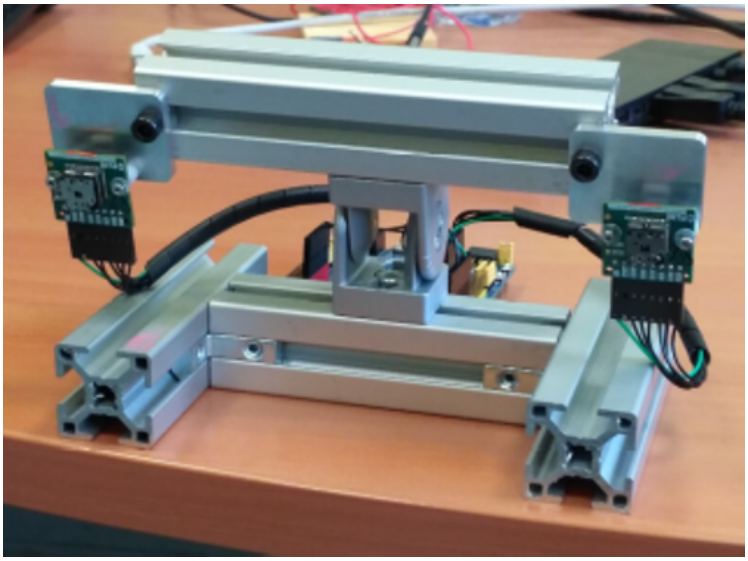


Fig. 1. The stereo system composed of two Lepton 2 cameras set in parallel

.

In order to detect falls, we decided to combine two thermal cameras in stereo mode. But if thermal cameras with a good resolution are relatively expensive, recently, some manufacturers have proposed very low cost thermal cameras, such as the Lepton 2 from FLIR, which we chose in our study. On the other hand, these low-cost cameras have a very low resolution (80×60 pixels), produce noisy images with drifts in the level of values over time (the cameras correct this drift from time to

Dimension	8.5 x 11.7 x 5.6 mm
Resolution	80 (h) x 60 (v) pixels
Pixel size	17 μm
Field of view	51° (h) x 37.83° (v)
Thermal sensitivity	<50 mK
Accuracy	± 5°C
Frame rate	9 Hz
Dynamic range	-10 to 140°C
Price	<200$

Table 1: Characteristics of Lepton 2.5

time, which has the effect of an abrupt temporal jump in the values of the image) and do not have a temperature calibration.

The acquisition system is composed of a pair of FLIR lepton 2.5 cameras (Table 1). The horizontal field of view is 51° and the diagonal field of view is 63.551°. The maximum frame rate of the cameras is eight frames per second. The two cameras are not synchronized. They are placed in parallel (their optical axes are parallel) in order to favor the field of view. The distance between the two cameras (the base line) was defined at 16 cm. The device will be placed high up (on the ceiling or on a wall just below the ceiling) and directed so as to monitor an entire room.

The two cameras are controlled by a micro-controller card. This card allows to recover the images on a PC by USB and in the future to make the interface with the embedded processing card.

3 Camera calibration

Conventionally, the calibration step consists in finding the parameters of a calibration model from real points given by a calibration grid. In our case, we chose the pinhole camera model and we used the functions of the OpenCV library to estimate the different parameters of the model. However, we were faced with two problems: 1) making a calibration grid suitable for the physical properties measured by the sensor and its resolution and 2) ensuring that despite the low resolution of our cameras, the process calibration is robust, i.e. precise and reproducible.

3.1 Calibration grid

The calibration grid must give temperature information. As the calibration process can be quite long, we opted to manufacture a grid whose points are actively heated. We have placed 36 bulbs, in

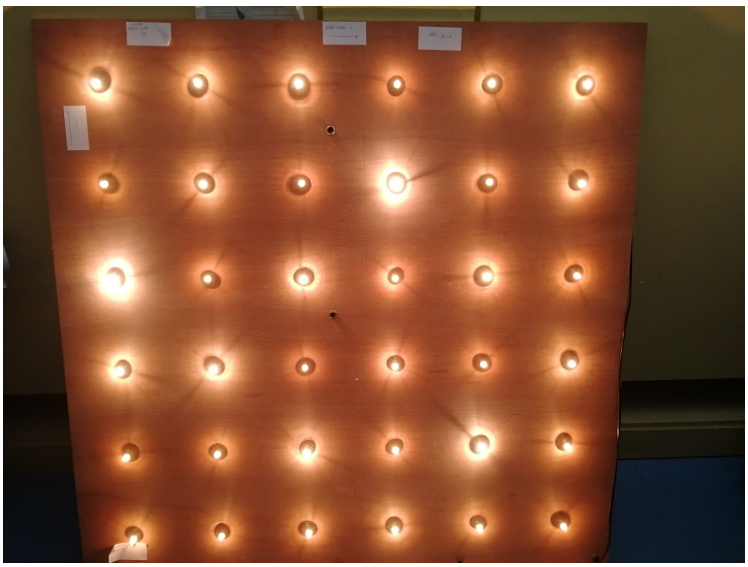


Fig. 2. Calibration grid

the form of a 6×6 grid, on a wooden panel (Fig 2). Each bulb is separated from its neighbor by a distance of 160 mm, which gives us a panel of $1\,\text{m} \times 1\,\text{m}$.

3.2 Robust calibration

The robustness of the calibration will depend, on the one hand, on the precision of the estimation of the points of the grid in the images and, on the other hand, on the number of different acquisitions of this grid.

Due to the low resolution, we replaced the pixel-precise point extraction module offered in OpenCV by our own method where the center of the light halos is located to sub-pixel precision. The other parameter that will influence the accuracy of the calibration is the number of grid images used to estimate the model parameters. Too small a number not only results in an inaccurate but also non-reproducible calibration. In this case, the parameters estimated by two successive calibrations can be quite different. On the other hand, too large a number of images considerably increases the handling during the acquisition of the grid. We carried out a bootstrap-type study in order to assess the evolution of the precision and reproducibility of the calibration as a function of the number of images. We concluded that for our protocol, the acquisition of 35 image pairs was necessary and sufficient to provide an accurate and robust calibration.

4 Features extraction and stereo vision

4.1 Extraction and matching of feature points in subpixel precision

Once the cameras have been calibrated, 3D vision requires 3 steps:

1. the features extraction from the two images
2. the features matching to estimate the disparity between the two images
3. the triangulation to estimate the 3D position as a function of the disparity.

A good reconstruction then depends on the number of features extracted from the images and on the precision of the estimate of the disparity between the points seen by the two images. In order to maximize these two criteria, we have proposed the following diagram (Fig. 3).

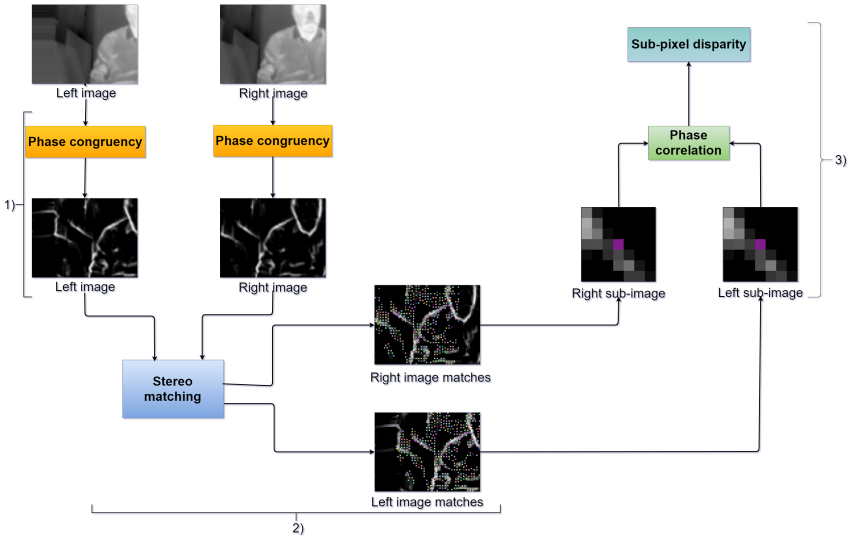


Fig. 3. Suite of treatments for matching to a sub-pixel precision: 1) robust features extraction; 2) robust matching to pixel precision; 3) more accurate estimation of the disparity at subpixel precision.

4.2 Features extraction

As we mentioned previously, thermal images are very little textured compared to images from cameras working in the visible spectrum. It is therefore much more difficult to extract a large number of landmarks (corners, edges) from these images. In addition, the low resolution makes these landmarks much more blurry, which leads to a much less precise localization. The literature gives us several classical methods for RGB images: the Harris corner detector [1], KLT [2], FAST [3],

BRIEF [4], the phase congruency [5, 6]. Some authors have proposed to adapt or simplify these methods for the extraction of landmarks in the case of thermal images (but on images of greater resolution). Hajebi et al. demonstrated in their paper that phase congruency could extract more characteristic points than other [7] methods. We therefore explored and adapted this method to our images.

In [8], the authors modified a bit the equation to circumvent the previous version drawbacks:

$$PhaseCong_2(x) = W(x)\frac{\lfloor \sum_n A_n[\Upsilon_n - \Psi_n] - T \rfloor}{\sum_n A_n(x) - \varepsilon} \tag{1}$$

where $\Upsilon_n = \left|\cos(\phi_n(x) - \bar{\phi(x)})\right|$; $\Psi_n = \sin(\phi_n(x) - \bar{\phi(x)})$, $W(x)$ is the frequency spread weighting, T is a noise threshold and ε is a small value (eg. 1-e3) to avoid division by 0.

The equation (1) can be extended to the two-dimensional image domain by applying it on several orientations θ after filtering the image by a bank of Log-Gabor filters. To reduce the computation cost due to the number of orientations and scales of the Log-Gabor filters bank, we implemented the variant proposed by [9] using a monogenic filter instead of Log-Gabor. The monogenic signal is a Riesz transform concatenated with a 2D signal.

As results of the 2D extension, we get a set of $PC(\theta)$, the phase congruency at orientation θ. The features are then estimated by combining all the $PC(\theta)$. This is done by computing the following values [8]:

$$a = \sum (PC(\theta)cos(\theta))^2 \tag{2}$$

$$b = 2\sum (PC(\theta)cos(\theta))(PC(\theta)sin(\theta)) \tag{3}$$

$$c = \sum (PC(\theta)sin(\theta))^2 \tag{4}$$

Combining a, b and c gives a hint of the strength of the feature. More precisely, the maximum moment M and the minimum moment m can be estimated by:

$$M = \frac{1}{2}(c + a + \sqrt[2]{b^2 + (a-c)^2}) \tag{5}$$

$$m = \frac{1}{2}(c + a - \sqrt[2]{b^2 + (a-c)^2}) \tag{6}$$

These moments are used to characterize the features. Higher is the maximum moment more significant will be the feature, and higher is the minimum moment more probably this feature point will be a corner. Because of the lack of information in images, we only took M into accounts. So a feature was considered significant if M was higher than a threshold γ.

In short, phase congruency reflects the behavior of the image in the frequency domain. It has been noted that contour or corner type elements have several of their frequency components in the same phase. The idea is then to seek the congruence of phases at different orientations and scales.

For each pixel, a maximum moment is estimated by combining the different congruences. A pixel having a maximum moment greater than a threshold γ will be considered as a feature.

An evaluation carried out on our low resolution images made it possible to demonstrate that: the phase congruency made it possible to extract more features than the other classical techniques and that the phase congruency was insensitive to sudden temporal variations in the intensity of the images . We also found that by varying the value of the threshold γ, we could obtain either features which are robust but few in number, or many features but at the expense of robustness.

4.3 Features matching

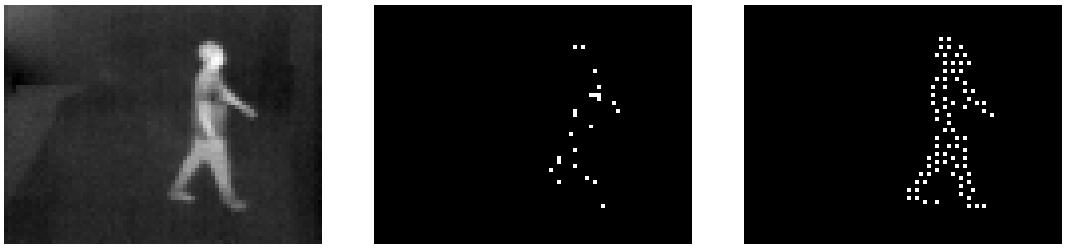
The calibration of the cameras makes possible to rectify the images. The matching is then simplified by finding similar points along the epipolar line. Similarity is estimated using the Lades similarity measure [10] conducted in a 5×5 window. The matching is then confirmed using consistency criteria between the images (constraints of uniqueness or order of the points along the line, similar orientation of the landmarks, bijectivity of the matching between right and left views, etc.). We visually assessed the suggested matches for 15 pairs of images and found less than 1% of incorrect matches.

4.4 Estimate the disparity at a subpixel level

Matching allows you to estimate the disparity for each pair of points to pixel precision. This precision is however insufficient because we showed that in the configuration of our assembly, a matching error of 1 pixel led to an uncertainty of more than 50 cm in depth. Phase correlation is a classic method of estimating a translation by finding the position of the maximum of the correlation peak. The idea is then to model this peak using a cardinal sine in our case and to estimate the position of the maximum at a sub-pixel level by fitting the model to the data. After estimating the best computational window size for the correlation, we evaluated our method on high resolution thermal images and on images acquired by our cameras. On high-resolution images, more than 99% of the dots were matched with an accuracy of less than 0.5 pixels. However, the match rate would decrease if higher precision was desired: 90% of matched pixels with an error less than 0.25 pixels, 55% with an error less than 0.1 pixels and 33% with an error less than 0.05 pixels. We observed identical behavior on our low-resolution images with match rates of 97%, 83%, 55% and 34% for errors less than 0.5, 0.25, 0.1 and 0.05 pixels, respectively.

4.5 3D reconstruction

As we had no ground truth we simply compared the 3D reconstruction of the points extracted, matched and reconstructed by triangulation using our method (phase congruency and phase correlation) with those reconstructed using more classical functions provided by OpenCV (features extraction using ORB [11], matching using KNN) (see Fig.4). We clearly see that our method reconstructs more 3D than the classical method. If we look at the distribution of the depth in z of the reconstructed 3D points for a person who is about 3 m from the cameras: ORB + KNN places



(a) Thermal image, left view

(b) Re-projected 3D points (ORB + KNN) (# 24)

(c) Re-projected 3D points (our method) (# 80)

Fig. 4. 3D points re-projected on an image plane.

the points at 1523.94 ± 1612.76 m while our method gives 3648.10 ± 256.43 m. Our measurement seems more precise and provides points with less dispersion.

We also applied this method to hot spots placed on the ground. The idea was then to determine the ground plane in order to detect falls by analyzing the distance between the points extracted from people and the ground plane. Unfortunately, even though our method gave more precise 3D points, the scatter of the points was too large to estimate the ground plane robustly.

5 Fall detection

Classical stereovision methods based on 3D reconstruction have not been able to reconstruct the ground plane, nor to estimate the pose of the person from our pairs of low-resolution thermal images. This, despite the fact that we had developed a robust calibration method, and pattern extraction and matching at a sub-pixel level of precision.

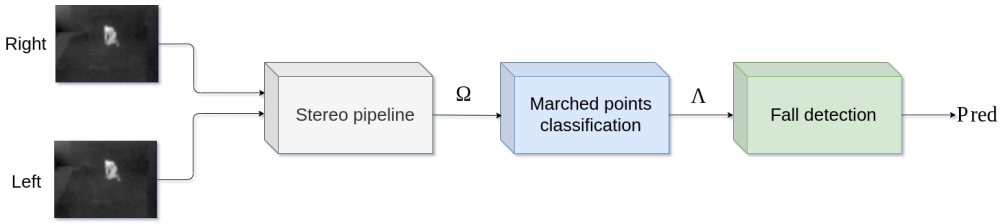


Fig. 5. Fall detection procedure

We then thought of another strategy: is it possible to define the relationship between a 3D point and its projections on the two images of the stereo pair, not analytically by a model, but by learning?

The idea is then to use a convolutional neural network or a more classic classifier and to teach it whether a 3D point seen by the two images of the pair is on the ground or not.

The fall detection procedure is as follows (Fig. 5):

1. A matching procedure. We have taken the procedure described in the chapter "Detection of falls by stereoscopic reconstruction" with:
   (a) features extraction by phase congruency;
   (b) A simple segmentation of these points to keep only the points potentially belonging to a silhouette. This segmentation is easy on thermal images because of the heat given off by people;
   (c) features matching by simple similarity measurement and some consistency criteria between the images. Note that, since we have not calibrated the cameras, we cannot use epipolar geometry. On the other hand, due to the lack of information in thermal images, there are ultimately relatively few points to match.
2. The classification of the paired points in "3D point on the ground" / "3D point above the ground" by inference of the network or the classifier;
3. The detection of the fall by an analysis of classified points.

It should be noted that we do a fall detection based on the analysis of static images. We do not involve a temporal analysis at this level.

The key points of the procedure are then the choice of the network or the classifier, the learning of this network / classifier and the analysis of the classified points.

5.1 Dataset

The objective is to find out if a pair of matched points is on the ground or above. For this, we moved a hot spot (a lit bulb) on the floor and in the room space above the floor and we acquired the corresponding images (Fig. 6). The lamp is easy to segment as it is usually the 'brightest' object. We therefore have a first learning base composed of pairs of images of the lamp associated either with a class " 3D point on the ground " or with a class " 3D point above the ground ".

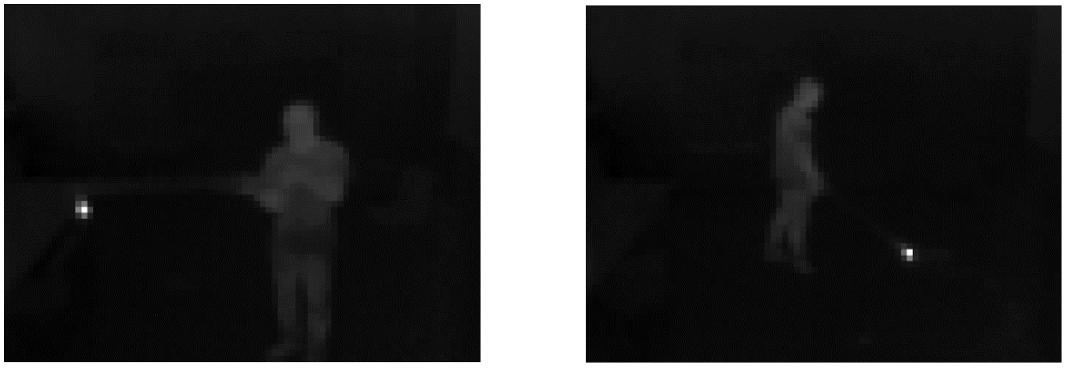
In order to simplify the learning, we also extracted the 2D position of the bulbs in the images (center of gravity in gray level in an ROI around the bulbs). We therefore have a second learning base made up of pairs of 2D coordinates associated either with a class "3D point on the ground" or with a class "3D point above the ground".

Our learning strategy allowed us to have a balanced data set containing as many points on the ground as there are non-ground points.

5.2 Choice of network or classifier.

We have explored two strategies depending on the information on features points: 2D coordinate pair or thermal image pair.

If 2D points are presented as their 2D coordinates in images, a simple SVM type classifier is sufficient. If the 2D points are in the form of thermal images, we have proposed a model based on deep learning, inspired by DenseNet (DGD) [12]. This solution has the advantage of associating the characteristics of the thermal image (resolution, noise, halo, etc.) in the learning process.



(a) Above ground **(b)** On the floor

Fig. 6. Left view of the image of a lamp

5.3 Analysis of classified points.

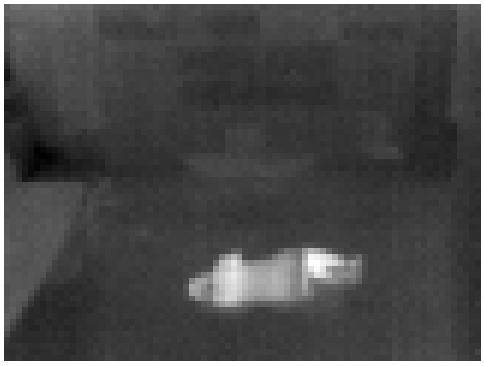
For a person on the ground, we assume that a large portion of the outstanding points will be classified 'on the ground'. Likewise, for a person standing, sitting, or even lying in a bed or on a sofa, only a small part of the features points (those of the feet for example) should be classified on the ground. A simple threshold on the percentage of points classified on the ground should be sufficient to detect or not the fall.

5.4 Results

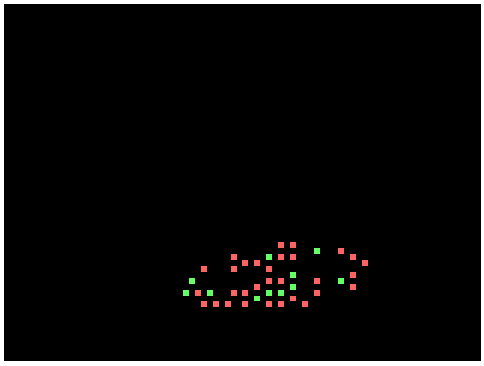
For training, we acquired 6000 images of the lamp on the ground and 6000 images of the lamp above the ground. This acquisition is done in less than half an hour because it suffices to move the lamp on the ground or in space with the cameras in video mode (8 fps). Such an acquisition must be done once for a given configuration of cameras.

The performance of our fall detector has been tested on four databases. Each database was made up of images acquired from a different person. In these images the people were in one of the following configurations: standing, walking, sitting, lying on a bed or a sofa or fallen on the floor. These databases were not balanced in the sense that there were relatively few drops compared to other activities.

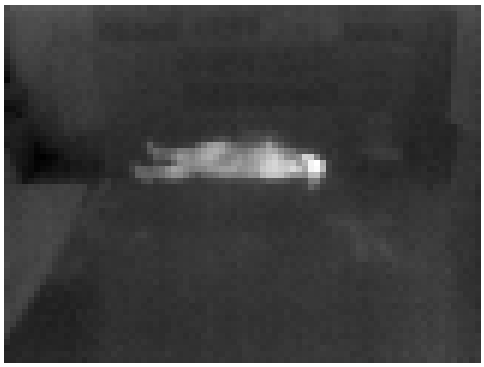
First, we compared our two classifier strategies: SVM on 2D coordinates and DGD on image data. For this we applied our classifiers to the training data in the form of 5 replications of a 2-block cross-validation (5x2cv). For each replication, we randomly shuffled the data. However, we have learned and tested SVM and DGD with the same data respectively. The median classification rate of DGD (0.976) slightly exceeds that of SVM (0.965) ($p = 0.00028$). Taking the image information into account provides a slight gain in classification precision. However, this gain seems a little marginal compared to the complexity of the implementation of the DGD.



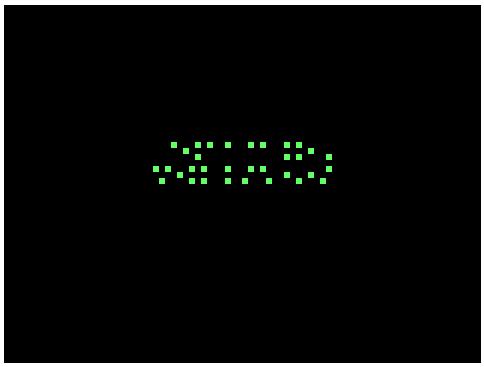
(a) Person after falling to the ground.



(b) 76.66 % points classified on the ground.



(c) Person lying on a sofa



(d) 0% points classified on the ground.

Fig. 7. Difference between a person on the floor (a and b) and a person lying on a sofa (c and d)

We then set the parameters of our method (the phase congruency threshold which acts on the number of features and the decision threshold on classified points) on data of persons. As we knew the ground truth fall/no fall , we were able to evaluate and optimize the different settings in terms of sensitivity, specificity, precision and F1-score. The results obtained showed very good performance with scores greater than 0.9 for the various indicators.

On the other hand, we found that our method was able to distinguish a person lying down, a person on the ground. This case is generally difficult to discriminate for single-camera methods, even based on deep learning (Fig. 7).

6 Conclusion

In this paper, we presented a framework to perform fall detection for very low-resolution thermal images. The framework is composed of three main parts: Points classification, Stereo-pipeline and a threshold-based fall detection. We prove that phase congruency is not only adapted to feature extraction for thermal image but also that stereo matching performed on phase congruency magnitude space provide more and well distributed matches than a well-traditional method that is ORB.

We compared a linear SVM method with a network witch we had derived from DenseNet. By classifying if a point is on the ground or not, we could determine the percentage of an object which is on the ground. So by using an appropriate threshold, we could perform fall detection.

Acknowledgments

This work was part of the PRuDENCE project (ANR-16-CE19-0015) which has been supported by the French National Research Agency (ANR).

References

[1] Harris C, Stephens M. A Combined Corner and Edge Detector. In: Procedings of the Alvey Vision Conference 1988; 1988. p. 147–151.

[2] Tomasi C, Kanade T. Shape and motion from image streams: a factorization method. Proceedings of the National Academy of Sciences. 1993;90(21):9795–9802.

[3] Rosten E, Porter R, Drummond T. Faster and better: a machine learning approach to corner detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2008;32(1):105–119.

[4] Calonder M, Lepetit V, Strecha C, Fua P. BRIEF: Binary Robust Independent Elementary Features. In: Computer Vision ECCV 2010. Springer Berlin Heidelberg; 2010. p. 778–792.

[5] Morrone MC, Owens RA. Feature detection from local energy. Pattern Recognition Letters. 1987;6(5):303–313.

[6] Kovesi P. Image Features from Phase Congruency. Videre J Comput Vision Res. 1999;1(3):C3–C3.

[7] Hajebi K, Zelek JSJS. Sparse disparity map from uncalibrated infrared stereo images. In: 3rd Canadian Conference on Computer and Robot Vision (CRV); 2006. p. 17–17.

[8] Kovesi P. Phase Congruency Detects Corners and Edges. In: Digital Image Computing: Techniques and Applications 2003. vol. 1; 2003. p. 309–318.

[9] Wang L, Zhang C, Liu Z, Sun B, Tian H. Image feature detection based on phase congruency by Monogenic filters. In: The 26th Chinese Control and Decision Conference (2014 CCDC); 2014. p. 2033–2038.

[10] Lades M, Vorbruggen JC, Buhmann J, Lange J, von der Malsburg C, Wurtz RP, et al. Distortion invariant object recognition in the dynamic link architecture. IEEE Transactions on Computers. 1993;42(3):300–311.

[11] Karami E, Prasad S, Shehata M. Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images. arXiv preprint arXiv:171002726. 2017.

[12] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–4708.