A Hierarchical Approach for Indoor Action Recognition from New Infrared Sensor Preserving Anonymity

Félix Polla[1], Hélène Laurent[2] and Bruno Emile[1]

[1]*University of Orleans, Prisme Laboratory EA 4229, Orléans, France*

[2]*INSA CVL, University of Orleans, Prisme Laboratory EA 4229, Bourges, France*

Keywords: Low Resolution Infrared Sensor, Motion History Image (MHI), Feature Selection, Action Recognition.

Abstract: This article is made in the context of action recognition from infrared video footage for indoor installations. The sensor we use has some peculiarities that make the acquired images very different from those of the visible imagery. It is developed within the CoCAPS project in which our work takes place. In this context, we propose a hierarchical model that takes an image set as input, segments it, constructs the corresponding motion history image (MHI), extracts and selects characteristics that are then used by three classifiers for activity recognition purposes. The proposed model presents promising results, notably compared to other models extracted from deep learning literature. The dataset, designed for the CoCAPS project in collaboration with industrial partners, targets office situations. Seven action classes are concerned, namely: no action, restlessness, sitting down, standing up, turning on a seat, slow walking, fast walking.

1 INTRODUCTION

In recent years, automatic recognition of human activities has attracted a lot of attention in the field of computer vision. Action recognition can be used for analyzing the behavior of people or for monitoring living and workplace environments (Jalal et al., 2017; Laptev et al., 2008). Despite impressive performances, the main pitfall faced by the techniques developed for home automation is a marked restraint of the users to be filmed. This led to feasibilty studies for human activity recognition through unobtrusive sensors such as pyroelectric infrared sensors (Luo et al., 2017), bluetooth low-energy beacons (Filippoupolitis et al., 2017) or acceleration sensors incorporated in smartphones or smartwatches (Sefen et al., 2016). To cope with the problems of personal identity revealed in operation, one of the objectives of the CoCAPS project is to consider the feasibility of action recognition using images coming from a low resolution (64×64 pixels) infrared sensor which guarantees the respect of the intimacy of the person. The sensor developed by Irlynx[1] within the project is a prototype able to return images of moving objects in a room. The technology is based on the principle of pyroelectric detection, that is to say, the detection of displacement of a hot body present in the monitored volume. It allows to observe bright or low-light scenes that would be otherwise difficult to monitor. This type of sensor, up-and-coming, is not yet completely mastered. The main disadvantage is that the provided images present a very noisy aspect, which brings up a problem of image quality (see Figure 1). Arranged in the ceiling upon request from building industrialists, we have then to face two difficulties: the considered sensor provides unusual data and unusual field of vision, so that no public database nor work dealing with this problem are available.

In this article, we detail the database constructed for the CoCAPS project, the proposed hierarchical approach for action recognition and its performances that achieve competitiveness with models extracted from deep learning literature.

2 DATA REPRESENTATION AND FEATURE EXTRACTION

Activity recognition often breaks down into three main phases: segmentation, data representation and feature extraction, and finally classification. Representation and feature extraction have crucial influence on the performance of recognition, therefore it is es-

[1] www.irlynx.com

Polla, F., Laurent, H. and Emile, B.
A Hierarchical Approach for Indoor Action Recognition from New Infrared Sensor Preserving Anonymity.
DOI: 10.5220/0008942002290236
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 229-236
ISBN: 978-989-758-402-2; ISSN: 2184-4321

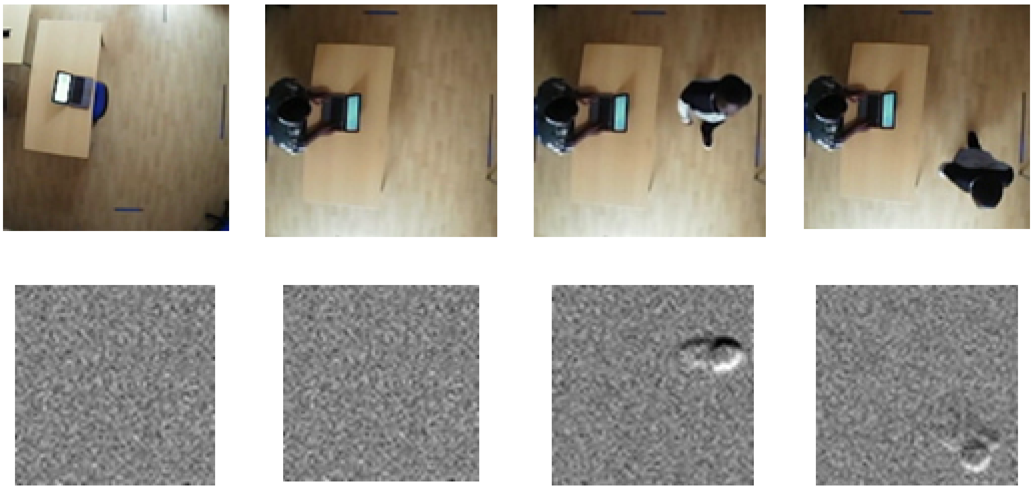


Figure 1: The first row presents images with visible camera respectively without and with the presence of a moving person in the scene; the second row presents the corresponding images from the infrared sensor.

sential to extract or represent features of image frames in a proper way. In this article, we describe some approaches used to represent a video sequence in order to combine them and extract representative information for each video.

2.1 Representation of Video Sequence

In 2D image sequences, many representations are possible such as silhouettes (Bobick and Davis, 2001; Gorelick et al., 2007), key pose (Liu et al., 2013), optical flow (Robertson and Reid, 2005), trajectory description (Wang et al., 2013), local descriptors: SIFT (Lowe, 2004), HOG (Dalal and Triggs, 2005).

Given the main constraints related to the used sensor and its usage context which are: noisy images, no detection on images if there is no movement of a hot body and sensor positioning in top view, we opt for an approach based on the silhouette. This approach implies that the segmentation and post-processing step is effective so that the subsequent processing can operate. In the work carried out previously (Polla et al., 2017), a method of segmentation and post-processing adapted to the sensor has been proposed.

Among the silhouette-based methods, some authors proposed the concept of Spatio-temporal template (Bobick and Davis, 2001). They first segment the images to extract the motion image, then for a set of images, they construct the motion history image (MHI) and the motion energy image. The MHI is a temporal model of video representation quite simple and widely used for the recognition of action (Ahad et al., 2012). From this information, it is possible to extract a large number of features.

In our approach, we use equation (1) to build the motion history image. It allows to group consecutive images of each video sequence according to their time information.

$$MHI(x,y,t) = \max_{i=\{1,..,m\}} \{D(x,y,t-(m-i))-10(m-i)\} \tag{1}$$

where $D(x,y,t)$ is the binary motion detection image extracted from the segmentation step at time t, m represents the selected time frame (m=15) and the value 10 is a threshold for varying the grayscale in the construction of the motion history.

2.2 Feature Extraction

In the literature, feature extraction approaches are classified into two families, namely contour-based methods and region-based methods (see Figure 2).

We propose a framework combining the contour-based method and the region-based method. In this section, we briefly present the three used descriptors.

2.2.1 Hu Moments

In this approach, central moments of any order are computed from raw moments and standardized to construct an invariant descriptor in translation and in scale. Finally, to also achieve invariance to the rotation, Hu (Hu, 1962) reformulated the above-mentioned moments by defining 7 new measures.

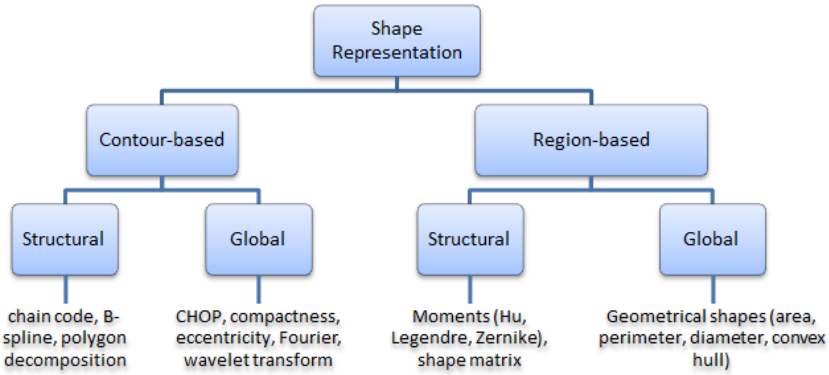

Shape Representation
Contour-based
Region-based
Structural
Global
Structural
Global
chain code, B-spline, polygon decomposition
CHOP, compactness, eccentricity, Fourier, wavelet transform
Moments (Hu, Legendre, Zernike), shape matrix
Geometrical shapes (area, perimeter, diameter, convex hull)

Figure 2: Numerous approaches for shape representations.

2.2.2 Color Histogram of Oriented Phase (CHOP)

The CHOP descriptor (Ragb and Asari, 2016) can accurately identify and locate image characteristics on gradient-based techniques. The features are formed by extracting the phase congruence information for each pixel, convolving the image with a pair of quadrature log-Gabor filters to extract the local frequencies and phase information.

2.2.3 Geometrical Shape

Several geometric indices have been proposed in the literature (Coster and Chermant, 1989). Some examples used for this study are : area (A), ellipse characteristics (eccentricity, major axis: R_{max}, minor axis: R_{min}), convex hull (C_H), perimeter (P), circularity ($\frac{R_{min}}{R_{max}}$), perimeter convexity ($\frac{P(C_H)}{P}$), surface convexity ($\frac{A(C_H)}{A}$).

2.3 Feature Selection

On a single image, a large number of data can be extracted. This can result in the known problem of the ”dimension curse” related to classification algorithms because considering a high number of attributes increases the risk of taking into account redundant or correlated ones which makes these algorithms more complex (storage space and high learning time) and sometimes less effective. In the literature, the feature selection approaches are divided into 2 categories: feature ranking algorithms (Biesiada et al., 2005) and subset selection algorithms (Yu and Liu, 2003; Hall, 1999). In this article we use 3 methods of feature selection from the literature to find the more adapted one for the considered situation.

CFS (Correlation-based Feature Selection) (Hall, 1999) is an algorithm that classifies subsets of entities according to a heuristic evaluation function based on correlation. The bias of the evaluation function is oriented towards subsets containing entities strongly correlated with the class and having low correlation between them. Non-relevant entities should be ignored as they have a low correlation with the class. Redundant features must be hidden because they will be highly correlated with one or more of the remaining features.

ILFS (Infinite Latent Feature Selection) (Roffo et al., 2017) is a robust probabilistic feature selection algorithm based on a probabilistic latent graph that performs the ranking step while considering all possible subsets of features, such as paths on the graph. The relevance of an attribute is modeled as a latent variable in a PLSA-inspired generation process (Hofmann, 1999) that allows the importance of a characteristic to be studied when it is added to a set of attributes.

MRMR (Minimum Redundancy and Maximum Relevance) (Peng et al., 2005) is a feature selection method that allows to select features that are mutually distant from each other while having a high correlation with the classification variable. A comparative study on measures to assess the best redundancy and relevance was presented in (Auffarth et al., 2010).

3 OVERVIEW OF THE PROPOSED METHOD

For the CoCAPS project, several case studies have been targeted for action recognition. One of which relates to office environments, another one to institutions for older people. In these contexts manufacturers are interested in various scenarios that lead to

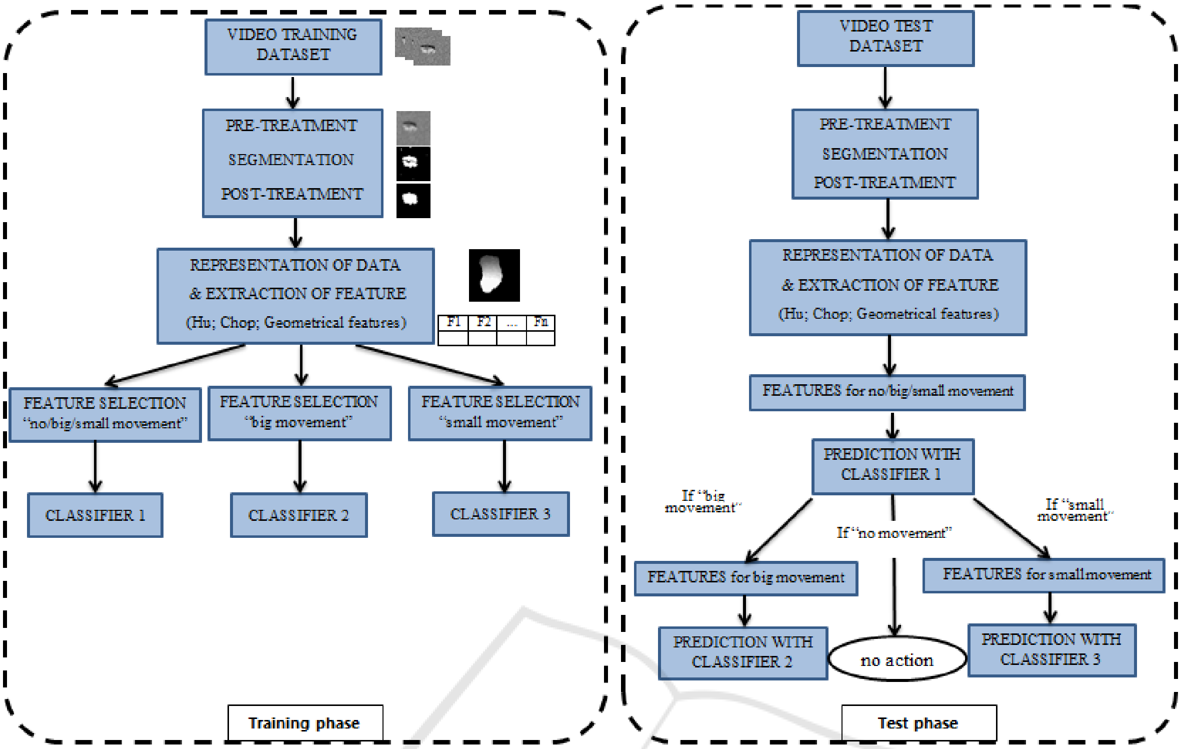

VIDEO TRAINING DATASET
PRE-TREATMENT
SEGMENTATION
POST-TREATMENT
REPRESENTATION OF DATA & EXTRACTION OF FEATURE (Hu; Chop; Geometrical features)
F1 F2 ... Fn
FEATURE SELECTION "no/big/small movement"
FEATURE SELECTION "big movement"
FEATURE SELECTION "small movement"
CLASSIFIER 1
CLASSIFIER 2
CLASSIFIER 3
Training phase
VIDEO TEST DATASET
PRE-TREATMENT
SEGMENTATION
POST-TREATMENT
REPRESENTATION OF DATA & EXTRACTION OF FEATURE (Hu; Chop; Geometrical features)
FEATURES for no/big/small movement
PREDICTION WITH CLASSIFIER 1
If "big movement"
If "no movement"
If "small movement"
FEATURES for big movement
FEATURES for small movement
PREDICTION WITH CLASSIFIER 2
no action
PREDICTION WITH CLASSIFIER 3
Test phase

Figure 3: Overview of the proposed approach.

the definition of fast walk / slow walk, big and small movements.

A movement is considered to be a big movement if there is a displacement of the center of gravity of the person and is considered as a small movement in the opposite case. Then in each group we find several actions. For example, in big movement we have fast walk (> 1 meter/second) and slow walk (< 1 meter/second). In small movement we have actions like: to sit down or get up, to stir (which includes for example typing, answering the phone, browsing documents) and to turn on a seat.

We propose a hierarchical model that exploits the Motion History Image and extracts a set of features from the MHI including geometric shapes, Hu moments and CHOP. In our approach, we define 3 classifiers: the first one to separate no movement, big movement and small one, the second one to separate the actions related to big movement and the last one for actions related to small movement.

Figure 3 shows an overview of the method. For each classifier, the definition of a dedicated features set is essential to have good performance. The left part of Figure 3 corresponds to the training phase of the model, within which we tested different feature sets including different combinations. During the test phase, we do not make a selection of features, but exploit the descriptors identified as being the best.

4 EXPERIMENTATION AND RESULTS

4.1 Dataset

We perform our experiments on the data acquired for the CoCAPS project. The dataset shows common actions in offices. It has 7 classes of human action: no action, restlessness, sitting, standing, turning on a seat, slow walking (speed less than 1 meter/second) and fast walking (speed greater than 1 meter/second). In total, the dataset consists of 700 videos samples (100 samples per action). The videos were taken considering various players and were collected at different times of the year. We consider clip sizes of 15 frames (about 1.5 second) for action recognition. In Figure 4, we present some results of the motion history image obtained for each action.

4.2 Protocol of Validation

We propose to conduct the performance study while using KNN (k-nearest neighbors) classification which

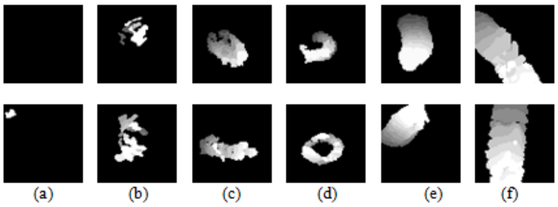


Figure 4: Example of motion history images for no action (a), restlessness (b), sitting or standing (c), turning on seat (d), slow walking (e) and fast walking (f).

is commonly used in machine learning. KNN is a classification method based on the closest training samples. To estimate the class associated with a new input, KNN algorithm consists in taking into account (in an identical way) the k learning samples which are the closest to the new input, according to a distance to be defined. There are different methods for comparing these values, like Hamming distance, Mahalanobis distance, Euclidean distance, etc. After various testings we used for our model the Manhattan distance, also known as city block distance, and k=3 neighbors. The Manhattan distance is calculated by making the sum of absolute differences between the coordinates of a pair of objects. This distance produces results close to those obtained by the simple Euclidean distance. However, with this measure, the effect of only one significant difference (outlier) is attenuated (Mohibullah et al., 2015).

To validate our results we use the K-fold validation (K=10) because it avoids the over-learning of the designed model. The K-Fold or cross-validation is a protocol in which individuals are separated into K groups of identical sizes. The learning takes place on K-1 groups and the validation on the withdrawn group. This operation is repeated for all groups and the average recognition rate is calculated. For model accuracy measure we use the F-score (Equation (2)).

$$F\text{-}score = 2.\frac{precision.recall}{precision + recall} \qquad (2)$$

where precision is the number of true positive results divided by the number of all positive results returned by the classifier and recall is the number of correct positive results divided by the number of all relevant samples.

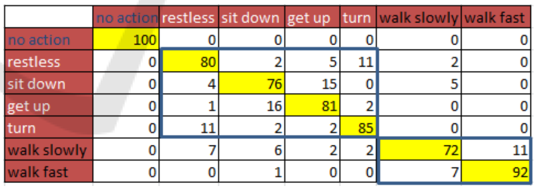
4.3 First Results and Interpretations

Before presenting the different results of the proposed hierarchical model, we first present those of the simpliest model (see Table 1) consisting in the raw classification into 7 classes (no action, restlessness, sitting, standing, turning on a seat, slow walking and fast walking) obtained with each type of descriptors or each combination of descriptors. The values of the f-scores do not exceed 83%, achieved when we just use the CHOP descriptor as features. Table 2 shows the confusion matrix obtained in that most favourable case. One can note that slow walking, within big movement, is not always appropriately recognized. It is mainly confused with fast walking and, to a lesser extent, with restlesness and sitting. Sitting is also confused with standing and inversely. As the classes of small and big movements are not very similar and also because it meets the needs of industrialists, we decided to test a hierarchical classification approach to first differentiate between no/big/small movement (classifier 1) and then classify within big movement (classifier 2) and small movement (classifier 3).

Table 1: Values of F-score for a model with a raw classification into 7 classes (single classifier).

Features	KNN *(7 classes)*
Hu (7 features)	49%
Geo (9 features)	64%
CHOP (128 features)	**83**%
Geo+Hu (16 features)	50%
Hu+CHOP (135 features)	71%
Geo+CHOP (137 features)	65%
Geo+Hu+CHOP (144 features)	52%

Table 2: Confusion matrix with CHOP as features: F-score=83%.



	no action	restless	sit down	get up	turn	walk slowly	walk fast
no action	100	0	0	0	0	0	0
restless	0	80	2	5	11	2	0
sit down	0	4	76	15	0	5	0
get up	0	1	16	81	2	0	0
turn	0	11	2	2	85	0	0
walk slowly	0	7	6	2	2	72	11
walk fast	0	0	1	0	0	7	92

We present in Table 3, the F-score values of the 3 classifiers with all combinations of the descriptors presented above. The first observation is that the separation of no, big and small movement (classifier 1) is easily done by using the geometrical shape descriptors or the combination of geometrical shape descriptors and CHOP with 97.8% of F-score.

The results of Table 3 also show that by merging descriptors, one can increase the performance of a classifier. This is the case of classifier 2 (big movement) and classifier 3 (small movement) where the fusion of CHOP and Hu moment descriptors allows to achieve 93.5% of F-score for classifier 2 and 82.5 % of F-score for classifier 3.

We tested the proposed hierarchical approach us-

Table 3: Classification rate of 3 classifiers with KNN.

Features	no/big/small movements *classifier 1*	big movements *classifier 2*	small movements *classifier 3*
Hu (7 features)	68.8%	90.5%	61%
Geo (9 features)	**97.8%**	86.5%	47.7%
CHOP (128 features)	97.1%	89.5%	81.5%
Geo+Hu (16 features)	84.4%	86.5%	47.7%
Hu+CHOP (135 features)	83.5%	**93.5%**	**82.5%**
Geo+CHOP (137 features)	**97.8%**	87%	51.5%
Geo+Hu+CHOP (144 features)	84.4%	87%	51.7%

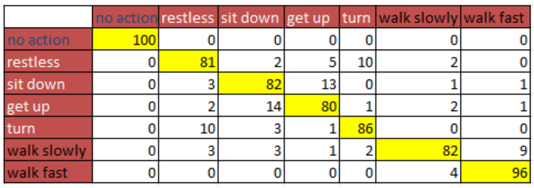
ing each time the most relevant descriptors for building each classifier. Two approaches were identified:

- Approach 1: using geometrical shape descriptors for classifier 1 and Hu moments + CHOP for classifier 2 and 3.
- Approach 2: using geometrical shape descriptors + CHOP for classifier 1 and Hu moments + CHOP for classifiers 2 and 3.

The hierarchical approach allows to achieve a F-score value of 86.5% for the first approach and 86.7% for the second approach.

Table 4 presents the confusion matrix of the proposed hierarchical approach 2. One can note that the recognition rate of quite all classes has been improved. The results are more uniform, especially for small movement. However confusions remain mostly between slow/fast walking, between sitting/standing and between turning on a seat/restlessness.

Table 4: Confusion matrix with hierarchical approach: F-score=86.7%.



	no action	restless	sit down	get up	turn	walk slowly	walk fast
no action	100	0	0	0	0	0	0
restless	0	81	2	5	10	2	0
sit down	0	3	82	13	0	1	1
get up	0	2	14	80	1	2	1
turn	0	10	3	1	86	0	0
walk slowly	0	3	3	1	2	82	9
walk fast	0	0	0	0	0	4	96

4.4 Results of Feature Selection

While combining different types of descriptors may improve the classification rate, having a large number of features for the model may conversely reduce its performance (calculation time and classification percentage). Ensuring a balance is necessary and the question of the choice of relevant features has to be dealt with. That is why we considered feature selections using three relevant methods from the literature, and analyzed the impact of the number of features for each method. These tests will provide the best features to constitute the classifiers.

To separate no, big and small movements, we use the combination of geometrical shape features and CHOP. In Figure 5 we note that by using the first 60 descriptors provided by the ILFS and MRMR methods, the F-score value is slightly increased.

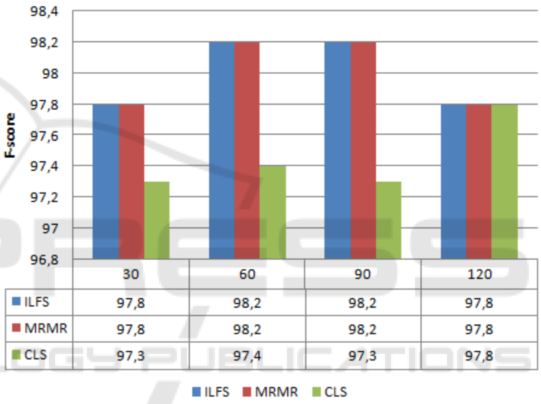


Figure 5: Results of feature selections for classifier 1: no, big, small movements.

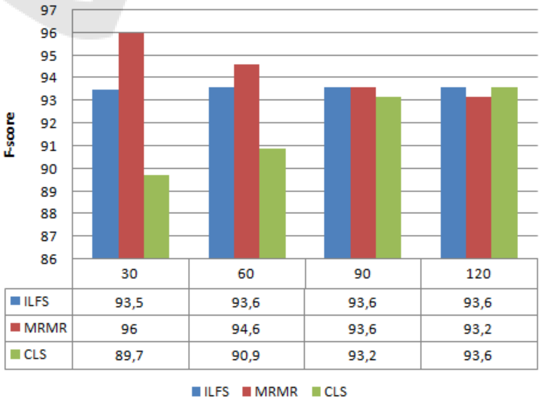


Figure 6: Results of feature selections for classifier 2: big movements.

Figure 6 shows that for classifier 2, by applying the MRMR selection method on the 135 descriptors resulting from the combination of Hu moments and

CHOP and using only the first 30 selected features, we can gain more than 2% of performance.

Concerning the selection of features for classifier 3 (small movement), Figure 7 shows that instead of the 135 descriptors resulting from the combination of Hu and CHOP, we can use 120 features selected by ILFS method while achieving a slightly increased performance.

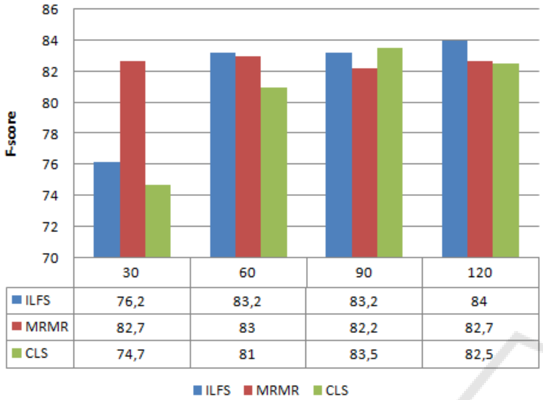
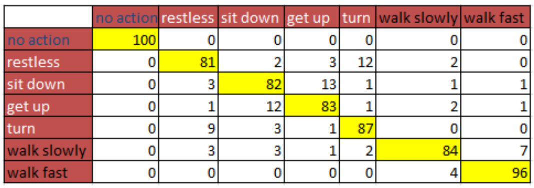


Figure 7: Results of feature selections for classifier 3: small movements.

Table 5 presents the confusion matrix of the proposed hierarchical approach using for classifier 1, the first 60 descriptors provided by the ILFS method, then for classifier 2, the first 30 descriptors provided by MRMR, and finally for classifier 3, the first 120 descriptors provided by the ILFS method. This selection allows to achieve a F-score value of 87.5%. It leads to an homogenization of the recognition rates within small movement but still fails discriminating slow and fast walking.

Table 5: Confusion matrix with feature selection: F-score=87.5%.



	no action	restless	sit down	get up	turn	walk slowly	walk fast
no action	100	0	0	0	0	0	0
restless	0	81	2	3	12	2	0
sit down	0	3	82	13	1	1	1
get up	0	1	12	83	1	2	1
turn	0	9	3	1	87	0	0
walk slowly	0	3	3	1	2	84	7
walk fast	0	0	0	0	0	4	96

4.5 Evaluation Review

We also compared the performance of the proposed hierarchical model with one depth learning approach, namely 3D-CNN (3D-Convolutional Neural Network), which is widely used for action recognition (Ji et al., 2012). In (Polla et al., 2019) it is shown that the 3D-CNN model is more efficient for our database than the LSTM model or 3D-CNN combined with the LSTM. This method aims at learning motion features by learning a hierarchy consisting of multiple layers of 3D spatio-temporal convolution kernels whose last layer output is used by a Multi Layer Perceptron (MLP) for classification.

Table 6: Summary table of comparisons.

Models	F-score *(7 classes)*
best result for raw classification	83%
3D-CNN with MLP	85%
hierarchical without selection	86.7%
hierarchical with selection	**87.5%**

Table 6 summarizes the different classification results obtained for the different tests. The deep learning model only achieve a F-score of 85%, which stresses the relevance of the proposed model.

5 CONCLUSION AND PERSPECTIVES

In this article, we present a hierarchical model for action recognition using low-resolution infrared video. This approach results in a classification rate of 87.5% while with a raw classification approach, 83% of F-score is achieved. The proposed approach even exceeds advanced methods such as the one based on neural networks. These results are quite interesting considering the type and quality of the sensor, constraint of the project (position of sensor in top view) and also the non-similarity of the shape within a same class (see Figure 4.(b) restlessness, for example). In future work, we plan to do action recognition on long videos. This will allow to answer the problematic of video sequence clipping for the classification.

ACKNOWLEDGEMENTS

The authors wish to thank BPI France, the Regional Councils of Limousin and Rhône-Alpes with FEDER, Department Council of Isère and District town of Bourges, Bourges Plus, for providing the financial support to CoCAPs project. The CoCAPs project, from FUI N°20, is also supported by clusters S2E2, Minalogic.

REFERENCES

Ahad, M. A. R., Tan, J. K., Kim, H., and Ishikawa, S. (2012). Motion history image: its variants and applications. *Machine Vision and Applications*, 23(2):255–281.

Auffarth, B., López, M., and Cerquides, J. (2010). Comparison of redundancy and relevance measures for feature selection in tissue classification of ct images. In *Industrial Conference on Data Mining*, pages 248–262. Springer.

Biesiada, J., Duch, W., Kachel, A., Maczka, K., and Palucha, S. (2005). Feature ranking methods based on information entropy with parzen windows. In *International Conference on Research in Electrotechnology and Applied Informatics*, pages 1–9.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):257–267.

Coster, M. and Chermant, J.-L. (1989). Précis d'analyse d'images. Technical report, Presses du CNRS.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*.

Filippoupolitis, A., Oliff, W., Takand, B., and Loukas, G. (2017). Location-enhanced activity recognition in indoor environments using off the shelf smartwatch technology and ble beacons. *Sensors*, 17(6):1230.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, New Zealand, Department of Computer Science, Waikato University.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2):179–187.

Jalal, A., Kim, Y.-H., Kim, Y.-J., Kamal, S., and Kim, D. (2017). Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern recognition*, 61:295–308.

Ji, S., Xu, W., Yang, M., and Yu, K. (2012). 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies.

Liu, L., Shao, L., and Rockett, P. (2013). Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern recognition*, 46(7):1810–1818.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Luo, X., Guan, Q., Tan, H., Gao, L., Wang, Z., and Luo, X. (2017). Simultaneous indoor tracking and activity recognition using pyroelectric infrared sensors. *Sensors (Basel)*, 17(8):1738.

Mohibullah, M., Hossain, M. Z., and Hasan, M. (2015). Comparison of euclidean distance function and manhattan distance function using k-mediods. *International Journal of Computer Science and Information Security*, 13(10):61.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8):1226–1238.

Polla, F., Boudjelaba, K., Emile, B., and Laurent, H. (2017). Proposal of segmentation method adapted to the infrared sensor. In *International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pages 639–650. Springer.

Polla, F., Laurent, H., and Emile, B. (2019). Action recognition from low-resolution infrared sensor for indoor use: a comparative study between deep learning and classical approaches. In *20th IEEE International Conference on Mobile Data Management (MDM)*, pages 409–414.

Ragb, H. K. and Asari, V. K. (2016). Color and local phase based descriptor for human detection. In *2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS)*, pages 68–73.

Robertson, N. and Reid, I. (2005). Behaviour understanding in video: a combined method. In *Tenth IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 808–815.

Roffo, G., Melzi, S., Castellani, U., and Vinciarelli, A. (2017). Infinite latent feature selection: A probabilistic latent graph-based ranking approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1398–1406.

Sefen, B., Baumbach, S., Dengel, A., and Abdennadher, S. (2016). Human activity recognition using sensor data of smartphones and smartwatches. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART), Volume 2*, pages 488–493.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79.

Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML)*, pages 856–863.