A new technique based on 3D convolutional neural networks and filtering optical flow maps for action classification in infrared video

A. Khebli *, H. Meglouli *,
L. Bentabet **, M. Airouche **

**Electrification of Industrial Enterprises Laboratory, University of Boumerdes Algeria*
(e-mail: a.khebli@univ-boumerdes.dz, hmeglouli@yahoo.fr)
*** Visualization and Computational Topology Laboratory Bishop's University of Canada,*
(e-mail: lbentabe@ubishops.ca, m_airou@yahoo.fr)

Abstract: Human action in video sequences provides three-dimensional spatio-temporal signals that characterize both visual appearance and motion dynamics. The aim of this work is to recognize human action in infrared video by focusing mainly on dynamic information. We developed a new technique based on deep 3D convolutional neural networks (3D CNNs) that take optical flow maps as input. Our approach consists mainly of three parts: 1) computation of optical flow maps; 2) filtering of these maps, using an entropy measurement in order to increase the classification rate and reduce the run time by eliminating sequences that do not contain human action; and 3) classification using 3D CNN. The experimental results obtained by our approach on the InfAR dataset show considerable improvement in comparison with results obtained by existing models.

Keywords: Artificial neural networks ,Image classification , Infrared imaging , Machine learning

1. INTRODUCTION

Human action recognition in 3D video is a major problem in research areas such as video surveillance, human-machine interfaces, video indexing and video search (Beaudry et al., 2016; Cao et al., 2016; Gao et al 2016; Hongyang et al., 2017; Ijjina and Chalavadi, 2016; Ji et al., 2013; Liu et al., 2016; Pablos et al., 2016; Turaga, 2008). Human action is defined as a periodic activity performed by one person in a time interval. Different types of actions, such as walking, jumping, running, etc., can be identified in a video sequence (Aggarwal and Ryoo, 2011; Chaquet et al., 2013; Turaga, 2008). Considerable progress has been made in recent years in human action recognition. However, it remains a difficult challenge due to the strong variation of human actions (Aggarwal and Ryoo, 2011; Chéron et al., 2015; Feichtenhofer et al., 2016; Pablos et al., 2016; Sheikh et al., 2005; Yu et al., 2017).

Several techniques proposed in the field of 2D image processing have been extended and adapted to videos, for example, 3D-SIFT (Scovanner et al., 2007), Extended SURF (ESURF) (Cheon et al., 2016; Willems et al., 2008), HOG3D (Gao et al., 2016; Holte et al., 2012; Klaser et al., 2008), etc. In addition, the optical flow has been widely used as an additional information element. The optical flow is defined as the apparent motion of individual pixels on the image plane (Horn and Schunck (1981)). Optical flow is considered a good approximation of the physical movement projected on the 2D image plane. Most of the methods used to compute optical flow assume that the colour/intensity of a moving pixel is invariant between consecutive images (moving from one image to the next). These approaches have shown excellent performance when applied on a variety of datasets (Holte et al., 2012a, b; Munaro et al., 2013; Subramanian et al., 2014). The temporal component of video provides an additional and important cue for recognition compared to single-image classification, as a number of actions can be reliably recognized based on the motion information (Chéron et al., 2015; Ji et al., 2013; Simonyan and Zisserman, 2014).

Convolutional neural networks (CNNs) can be considered a multilayer perceptron (MLP). CNNs were inspired by Hubel and Wiesel's work on the visual cortex in mammals (Baccouche et al., 2012). The first CNN dates back to the 1980s with K. Fukushima's (Fukushima, 1980) work on the neocognitron. The CNN model introduced by (LeCun et al., 1990, 1998, 2004) is the most popular deep neural network in use by the computer vision community. CNN-based methods were developed for object recognition, such as face recognition (Khalajzadeh, 2014; Lawrence et al., 1997; Matsugu et al., 2003), evaluation of video quality (Ijjina and Chalavadi, 2016) and recognition of human action in visible videos (Ji et al., 2013; Karpathy et al., 2014). (Z. Liu et al., 2016; C. Cao et al., 2016) proposed a 3D CNN to directly learn spatio-temporal characteristics from raw depth image sequences. In general, the advantage of using a CNN comes from the fact that it can proceed with image classification without going through the classic step of feature extraction.

In this paper, we developed a new technique that uses a 3D CNN for human action recognition in infrared videos. The objective of this approach is to exploit the temporal information in the video by calculating the optical flow maps from the infrared images, and then applying a 3D CNN on these maps.

The main contributions of this work can be summarized as follows:

1. We propose a new algorithm based on an entropy measurement to filter out the optical flow maps that do not contain human action. This step increases the classification rate and reduces the run time.

2. Through experimental tests on the InfAR, we demonstrate that the filtering step in 1) combined with the proposed 3D CNN significantly improves the classification accuracy in comparison with existing techniques.

This paper is organized as follows: Section 2 reviews the related work. Section 3 presents the proposed approach and the details of 3D CNN model. The experimental results and discussions are described in section 4. Finally, section 5 concludes the paper.

2. RELATED WORK

Most of the recent work on human action recognition is based on the use of CNNs, which led to an improvement of classification accuracy in comparison with previous work (Girdhar and Ramanan, 2017). These techniques try to benefit from the temporal and the spatial information of each image in the video in order to recognize the human action. Initially, 2D-space-based CNNs were extended in the time by loading consecutive video frames in a manner that allows the first layer to extract the spatio-temporal features (Ji et al., 2013). (A. Karpathy et al., 2014) studied several approaches for temporal sampling, including early fusion, slow fusion and late fusion. Their architecture is not particularly sensitive to the temporal modelling, and they achieved similar levels of performance by a purely spatial network. They indicated that their model does not gain much from the temporal information (Feichtenhofer et al., 2016). (D. Tran et al., 2015) proposed an approach that learns a 3D CNN from a limited temporal support of 16 consecutive images, all with 3 × 3 × 3 filter kernels. They reported better performances than (Karpathy et al., 2014) by letting all the filters operate on space and time. However, their network is considerably deeper (Feichtenhofer et al., 2016; Ji et al., 2013; Karpathy, et al., 2014). Recently, other techniques proposed the use of optical flow for spatio-temporal modelling. For example, (Simonyan et al., 2014) proposed an architecture that combines two CNNs: one for spatial information applied on pixels, and one for temporal information applied on dense optical flow maps. (C. Gao et al., 2016) used a network-like architecture (Simonyan and Zisserman, 2014)) with two independent deep CNNs, named "OF-MHI-CNN" and "Flow-CNN," respectively, where the two networks contain five convolutional layers and three fully connected layers.

In this paper, we use the same 3D CNN architecture proposed in (Gao et al., 2016; Tran et al., 2015). We propose to evaluate the entropy of all 30 optical flow maps in order to filter out the frames that do not contain enough human action. The remaining frames are processed by a deep 3D CNN to learn temporal information about a set of human actions. This method provides significant improvement for action recognition on infrared videos in comparison with existing techniques (Gao et al., 2016).

3. PROPOSED APPROACH

Our approach consists of three steps, as illustrated in Fig. 1.
Step 1: Optical flow map computation (each map represents the displacement of each pixel between two successive images).
Step 2: Filtering optical flow maps that do not contain information using the Entropy computation.
Step 3: Action classification by 3D CNN.

The entropy H of an image is defined as:

$$H = -\sum_{k=1}^{L} P_r(x_k)\log_2 P_r(x_k) \quad (1)$$

where L is the number of gray levels and $P_r(x_k)$ is the probability associated with gray level k .

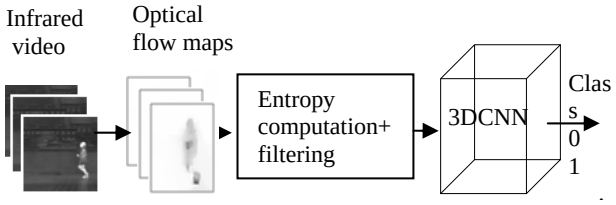

Fig. 1.General structure of the proposed approach.

3.1 Optical flow map computation

Optical flow is the pattern of apparent object motion between two consecutive frames caused by the movement of these objects or the camera. It is a 2D vector field where each vector is considered as a displacement vector showing the motion of points from the first frame to second (Lim et al., 2017; Liu et al., 2008). Optical flow assumes brightness constancy, which gives:

$$I(x,y,t) = I(x+dx, y+dy, t+dt) \quad (2)$$

Taking Taylor series approximation of right-hand side, and removing common terms and dividing by dt to get the following equation:

$$f_x u + f_y \vartheta + f_t = 0 \quad (3)$$

Where

$$f_x = \frac{\partial f}{\partial x} \quad ; f_y = \frac{\partial f}{\partial y} \quad (4)$$

$$u = \frac{dx}{dt} ; \vartheta = \frac{dy}{dt} \quad (5)$$

The above equations describe the optical flow in terms of the spatial image gradient with two unknowns (u, ϑ).f_x and f_y , are image gradients, and f_t is the gradient along time. This equation with two unknown variables cannot be solved. Various methods have been suggested to resolve this problem (Lim et al., 2017). We have based this work on the computation of optical flux given in (Liu et al., 2008). Fig. 2 shows the optical flow between pairs of consecutive frames.

a b c

Fig. 2. Optical flow (a), (b): Pair of consecutive frames; (c): Image of optical flow for (a) and (b).

3.2 Entropy computation and optical flow map filtering

In this section, we will focus on the entropy computation of each sequence of 30 optical flow maps and filtering sequences that do not contain information. Entropy computation steps are in algorithm 1.

Algorithm 1: Filtering optical flow sequences by entropy computation

Input: Sequences S {i} and their entropy E (i)
Output: **Remove *a* sequence S{i}**

1. **Q: Number of sequences in the video**
2. **For i=1: Q-1**
   $y(i) = |E(i+1) - E(i)|$ **(entropy difference computation between two consecutive sequences)**
3. if $y(i) \prec$ threshold
   Save (S{i},s{i+1})
   End if
   End for

Results of execution of this algorithm are reported in section 4.

3.3 3D convolutional neural networks

In this section, we present a short overview of 3D CNNs that will be used in our proposed model. 3D CNNs were proposed for human action recognition (Ji et al., 2013; Taylor et al., 2010). Most 3D CNNs process the data sequence as a fixed-frame sequence and apply 3D convolution layers for feature learning (He et al., 2017). Both spatial information and temporal information are abstracted layer by layer (Liu et al., 2017). The 3D pooling neural layers are used to reduce the number of parameters in the spatio-temporal lever. Usually, 3D CNNs have two different layers: the 3D convolution layer, and the 3D pooling layer. The function of the 3D convolution layer is to apply several convolutional filters over the input volumes. The 3D pooling layer selects the best feature extracted by the 3D convolution layer (He et al., 2017).

3.3.1 Notation

- $\mathbf{a}_{ij}^{x,y,z}$ denotes the value at position (x,y,z) on the j^{th} feature map in the i^{th} layer.

- $f()$ denotes the activation function, which is the $\tanh$ in the 3D convolution layer.

- m denotes the indexes over the set of feature maps in $th(i-1)$ layer connected to the current feature map j.

- b_{ij} denotes the bias term of the j^{th} unit of the layer i.

- P,Q,R represent the width and height of the image and the temporal dimension, respectively.

- $\mathbf{w}_{ijm}^{p,q,r}$ denotes the weight at position (p,q,r) of the kernel connected to the m^{th} feature map in the previous layer.

3.3.2 3D convolutional layer

In a 3D convolutional neural layer, there are many 3D convolution kernels; each is replicated over the image frame sequence. The 3D convolution layers play an important role in extracting spatio-temporal features, because they can encode temporal visual information from the video clips (He et al., 2017). We can compute the output of a 3D convolutional layer using the following formula:

$$\mathbf{a}_{ij}^{xyz} = f\left(b_{ij} + \sum_{m} \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} \sum_{r=0}^{R-1} \mathbf{w}_{ijm}^{pqr} \mathbf{a}_{(i-1)m}^{(x+p)(y+q)(z+r)} \right) \quad (6)$$

3.3.3 3D pooling layer

The pooling layer is another important operator in a CNN. A pooling operator runs on individual feature channels, coalescing nearby feature values into one via the application of a suitable operator. Common choices for this include max-pooling. The 3D pooling layer is computed as follows:

$$\mathbf{a}_{ij}^{xyz} = \max_{i,j,k \in \{1,2,\ldots,g\}} \mathbf{a}_{(i-1)j}^{(gx+i)(gy+j)(gz+z)} \quad (7)$$

Where g is the length of the pooling region.

3.4 3D CNN architecture

Fig. 3 presents the 3D CNN global architecture applied for human action recognition. This architecture comprises six layers, including two alternating layers of convolution and subsampling C1, S2 and C3, S4, followed by two layers of fully connected neurons, FC1 and Softmax. The size of the 3D input layer is 256 × 293 × 30, corresponding to 30 successive optical flow cards of 256 × 293 pixels each. The first layer is a convolution layer (C1) consisting of eight characteristic maps of 250 × 288 × 26 units. Each of them is connected with a 7 × 6 × 5 in the previous layer, called the "local receptive field." The next subsampling layer (S2) is composed of 08 size cards 125 × 144 × 26; each is connected to a feature card in C1. The convolution layer (C3) contains 16 feature maps of size 117 × 136 × 18 pixels, and each unit is connected to a 9 × 9 × 9 receptive field in the previous layer. The layer S4 follows the same principle as S2. Finally, the input information is encoded in a size vector 128 in the FC1 layer. This vector can be interpreted as a descriptor of the spatio-temporal information extracted from the input sequence. The Softmax layer contains a conventional multilayer perceptron with a neuron by a human action in the output layer. The detailed configuration is shown in Table 1. The first line refers respectively to the number of filters: the height, width and the temporal dimension of each filter. Stride specifies the intervals applied to the convolution kernel

in the input. Pad indicates the number of pixels to add to each side of the input.

4. EXPERIMENTS AND RESULTS

In this section, we discuss the performance of applying filtering in our technique on the InfAR dataset (Gao et al., 2016).

Secondly, we illustrate the influence of the using filtering algorithm on the classification rate of human actions by 3D CNN.

4.1 Dataset

We applied our technique based on 3D CNN model on InfAR dataset (Gao et al., 2016). This dataset consists of 12 classes of human action. Each of these classes is represented by 50 video clips of varying lengths, ranging from 80 images to 280 images. The frame rate is 25 frames per second, and the resolution is 256 x 293. Each video contains one person or several people performing one or multiple actions. For each class, we have 35 videos for training, and 15 videos for the test.

4.2 Results of our filtering algorithm

Fig. 4 shows the optical flow sequences of the jog action where each sequence contains 30 frames. According to the results obtained and illustrated in this figure, it was found visually that the sequences 1, 4 and 5 do not contain the action. In this case, these sequences are considered a noise and must be filtered to increase the accuracy of the algorithm based on 3D CNN.

Table 1. Configurations of global network.

C1	S2	C3	S4	FC1	softmax
8x7x6x5	Pool 2x2x1	16x9x9x9	Pool 4x4x4	128x1x1	12x1
Pad -	-	1x1x2	-	-	-
Dropout -	-	0.8	-	0.8	-
Stride 1	Stride 2	Stride 1	Stride 4	-	-

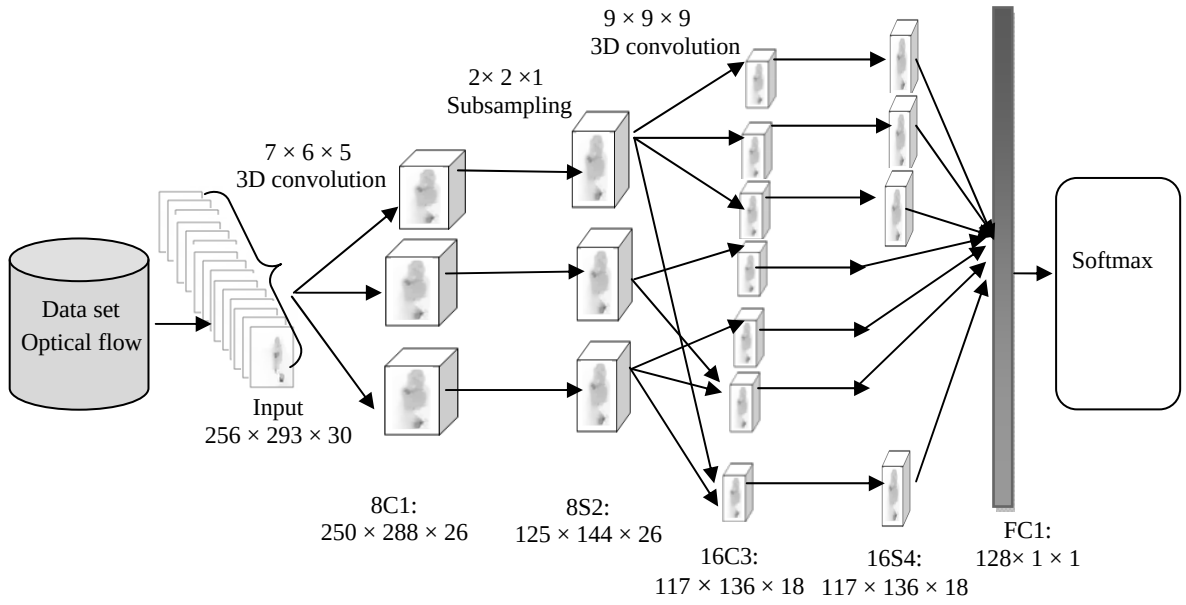

Fig. 3. 3D CNN global architecture applied for human action recognition.

Table 2 gives the values of the entropy $E(i)$, as well as the entropy difference between two consecutive sequences $y(i) = |E(i+1) - E(i)|$. This table also provides the saved sequences after filtering. First, we observe that $y(2) = |E(3) - E(2)| = 6.1579 \prec (\text{threshold})$. Our technique based on the filtering algorithm saves the sequences $s\{2\}, s\{3\}$ and deletes the other sequences automatically.

Fig. 5 shows the optical flow sequences of the multiple hand-wave action. From this figure, it was found visually that sequences 1 and 2 contain action.

Table 3 gives the values of the entropy $E(i)$. We can observe that $y(1) = |E(2) - E(1)| = 6.104 \prec 10(\text{threshold})$ and that applying our algorithm (algorithm 1) will save the two sequences $S\{1\}$ and $S\{2\}$.

4.3 Training and the validation results

Each iteration of the training using the 3D CNN model is performed by taking as input 90 optical flux sequences of length 30. The learning is performed using the stochastic gradient algorithm with momentum adapted to weight sharing. The momentum used is equal to 0.9, and the learning rate is equal to 10^{-5} at time t = 0. The learning rates must be divided by 2 to form the model for another 50 iterations. The learning is stopped when the learning rate is less than 5×10^{-6}. To reduce the over-fitting effect of the network, we use the dropout technique. For the evaluation, our technique

uses the Softmax function on the characteristic vector of the FC1 layer.

Table 2. Entropy and filtered sequences of the jog action.

Sequence (S)	Entropy (E)	y(i)	Threshold	Filtered sequences s{i}
s {1}	79.74	12.55	10	**s {1}, s {4} and s {5}**
s {2}	67.18	**6.16**		
s {3}	73.34	33.04		
s {4}	106.39	46.73		
s {5}	59.65			

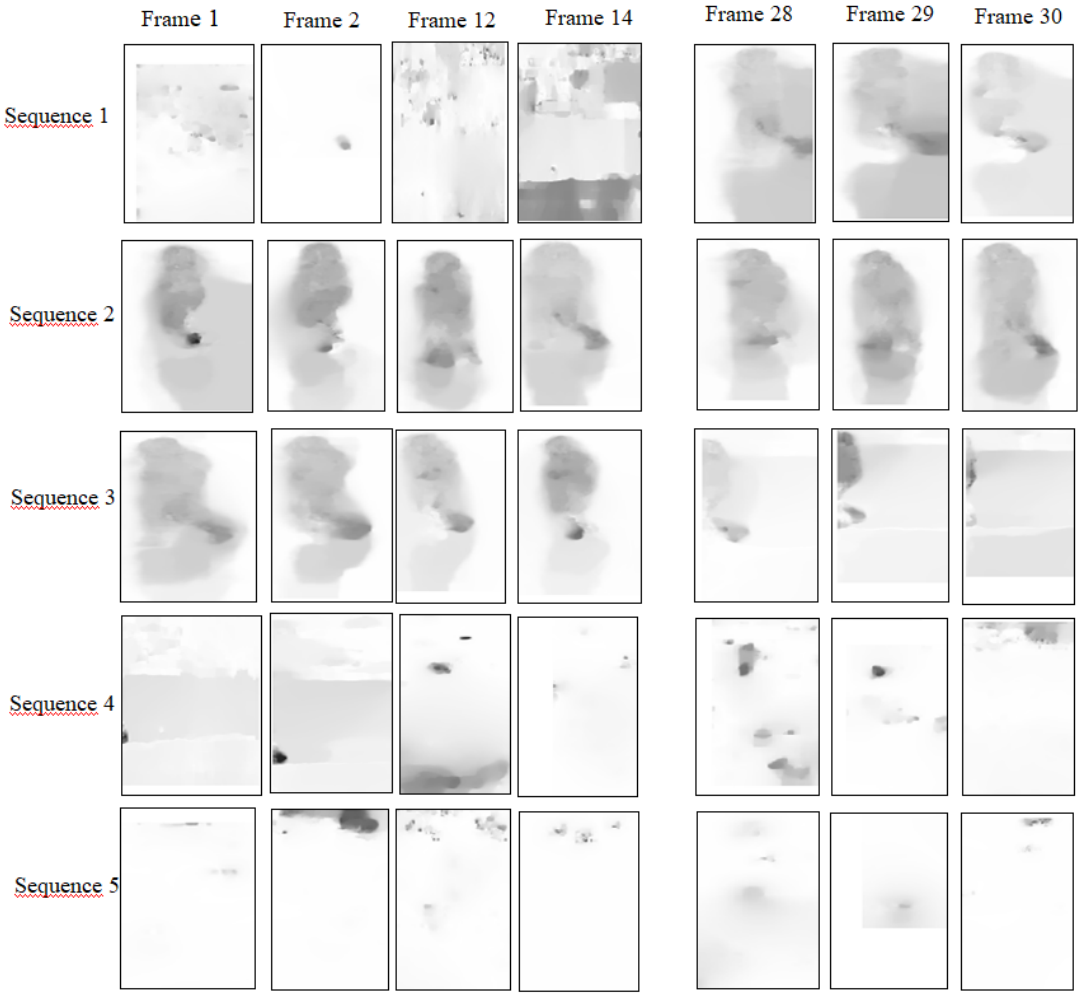

Fig. 4. Optical flow sequences of the jog action.

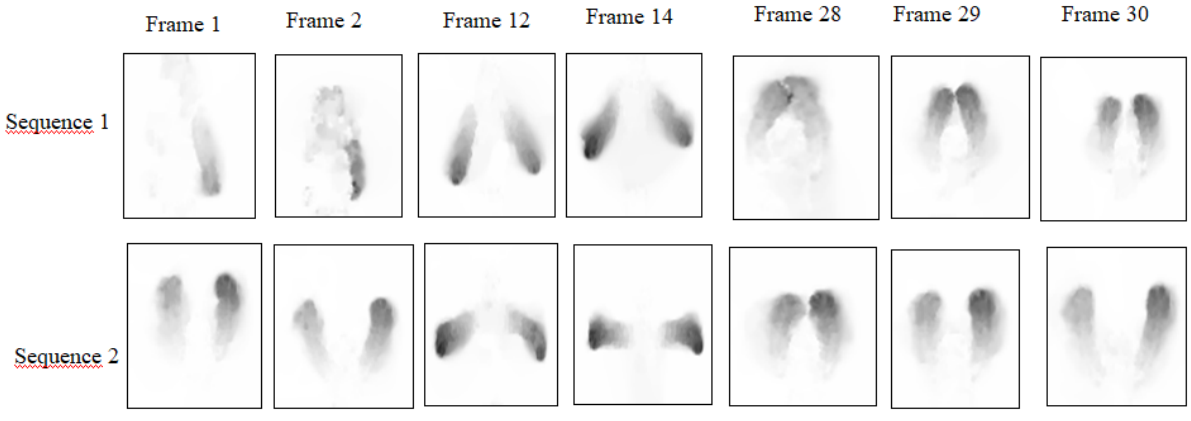

Fig. 5. Optical flow sequences of the multiple hand-wave action.

Table 3. Entropy and filtered sequences of the multiple hand-wave action.

Sequence (S)	Entropy (E)	y(i)	Threshold	Filtered sequences s{i}
s {1}	29.05	**6.10**	10	-
s {2}	35.15			

The first test of our technique is based on the action dataset used in (Chenqiang Gao et al., 2016). This dataset contains 12 types of human actions: one hand-wave (Wave 1), multiple hand-wave (Wave 2), handclap, walk, jog, jump, skip, handshake, hug, push, punch and fight.

To assign an action class to each test sequence, we performed the classification by 3D CNN, trained the action classifiers on the 1680 video sequences and tested the 60 video sequences. We then repeated this procedure for all the actions and calculated the average classification rate. Fig. 6 illustrates the confusion matrix obtained after using our filtering algorithm on the set of learning and test data. The classification rate is equal to 88.19% when the test and learning data are filtered by the filtering algorithm (algorithm 1) .

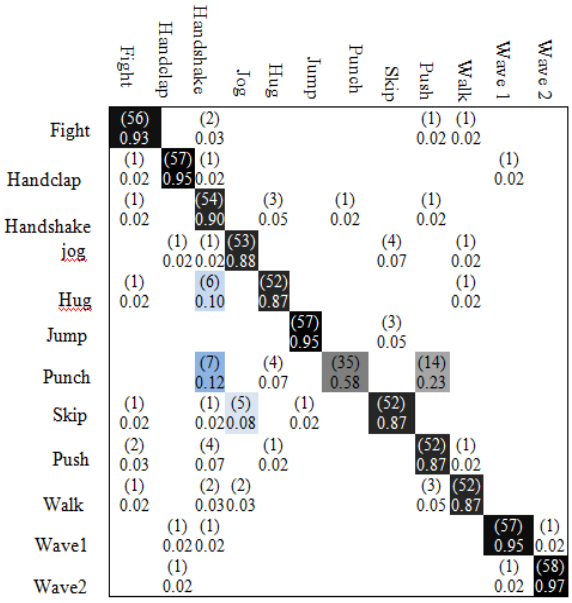

Fig. 6. Confusion matrix on the filtered InfAR dataset.

We observed that the punch action has a low accuracy, because this action is sometimes confused with the actions push, handshake and hug. The classification rate obtained in our work is considerably higher than that obtained in (Gao et al., 2016) , which was 76.66%. This improvement is due to the filtering of optical flow maps by our algorithm on the test and learning data (deletion of video sequences that do not contain the action).

In the confusion matrices shown in Fig. 7, we observed that when training and test data are not filtered, a classification rate of 53.88% is obtained. This rate is lower than the training data and the filtered test that are not filtered (Fig.8), which gives a classification rate of 77.5%. It is also lower than the test data that are filtered and learning that isn't filtered (Fig.9), which gives a classification rate of 73.61%. We also noted that the classification rate increased by 3.89%, reaching 77.5% (Fig.8) when the learning data is filtered compared to Fig.9, which represents the confusion matrix when the data of learning are not filtered.

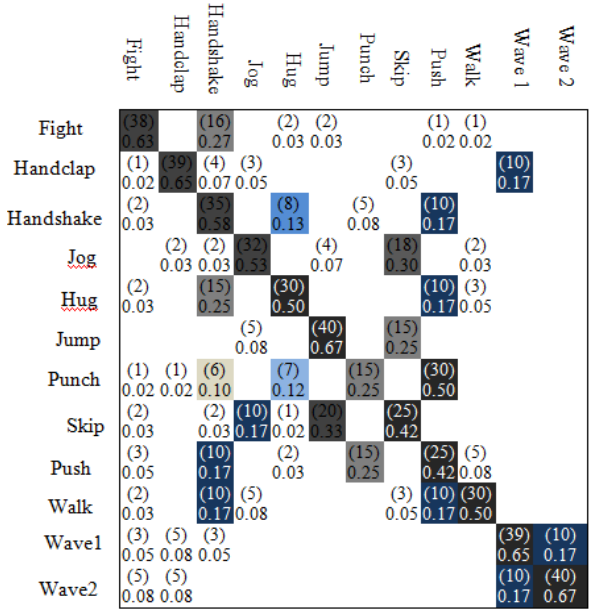

Fig. 7. Confusion matrix for unfiltered test and training data.

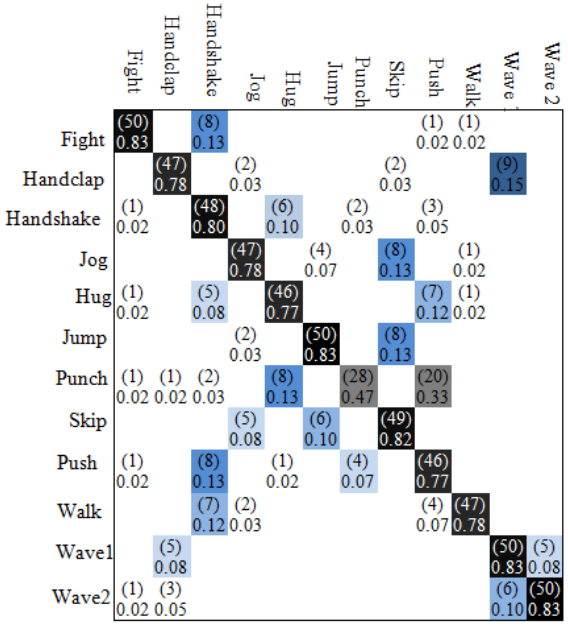

Fig. 8. Confusion matrix for the training data filtered.

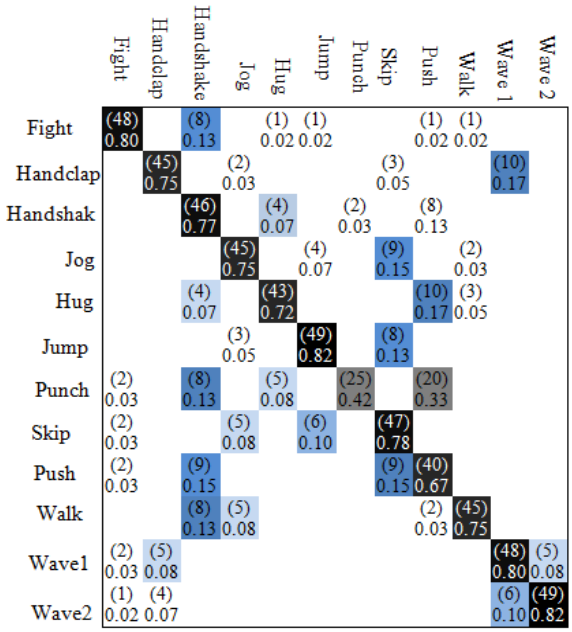

Fig. 9. Confusion matrix for filtered test data.

4.4 Comparative study

Table 4 shows a comparative study between our approach and state-of-the-art approaches applied on the same InfAR dataset (Gao et al., 2016). From Table 4, we observe that our method has the highest classification rate (88.19%), which is about 11.53 percentage points higher than Tow stream CNN (76.66%). The filtering of optical flow maps is important for the classification of video sequences.

Table 4. Comparisons of various methods.

Method	**Classification rate (%)**
STIP(Gao et al (2016))	49.16
3DSIFT(Gao et al (2016))	49.50
HOF (Gao et al (2016))	68.58
Two-stream without OF-MHI(Gao et al (2016))	32.08
Tow- stream CNN (Gao et al (2016))	76.66
Our method	**88.19**

5. CONCLUSION

In this paper, we have proposed a new method based on filtering optical flow maps and convolutional neural networks (3D CNNs) for human action classification. This technique exploits the motion information after filtering the maps of optical flow. The filtering method is based on computation of the entropy between two sequences of consecutive frames. Our technique is tested by applying the filtering algorithm on infrared video sequences to filter sequences that do not contain information (action). The results obtained by applying our filtering technique tested on different datasets shows considerable improvement. It allows increasing the ranking of all the actions in the InfAR dataset, reaching a classification rate of 88.19%.

REFERENCES

Aggarwal, J.K., Ryoo, M.S. (2011) . Human activity analysis: A review. *ACM Comput. Surv*. Vol . 43, no. 16, pp. 1–43.

Baccouche, M., et al (2010) . Action Classification in Soccer Videos with Long Short Term Memory Recurrent Neural Networks. Proc. *Inernational. Conference. Artificial Neural Networks – ICANN, Thessaloniki, Greece*, pp. 154–159.

Beaudry, C., Péteri, R., Mascarilla, L. (2016) . An efficient and sparse approach for large scale human action recognition in videos. *Machine Vision and Applications*., vol. 27, no. 4, pp. 529–543.

Cao, C., et al. (2016) .Action Recognition with Joints Pooled 3D Deep Convolutional Descriptors. Proc. *Inernational. Conference, Artificial Intelligence (IJCAI-16)*, New York, USA, pp. 3324–3330.

Chaquet, J.M., Carmona, E. J., Caballero, A. F.(2013). A survey of video datasets for human action and activity recognition.*Computer Vision and Image Understanding*, vol. 117, no. 6 , pp. 633–659.

Cheon, S.H., et al .(2016). An enhanced SURF algorithm based on new interest point detection procedure and fast computation technique. Proc. *Journal of Real Time Image Processing, Springer*, pp. 1-11.

Chéron, G., Laptev, I., Schmid, C .(2015). P-cnn: pose-based cnn features for action recognition. Proc. *Inernational. Conference, IEEE Computer Vision, Santiago, Chile*, pp. 3218–3226.

Feichtenhofer, C., Pinz, A., Zisserman, A .(2016). Convolutional two-stream network fusion for video action recognition. Proc. *Inernational. Conference. IEEE Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1933-1941.

Fukushima, K. (1980). Neocognitron : A self- organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, vol. 36 , no. 4, pp. 193–202.

Gao, C., et al .(2016). InfAR dataset: Infrared action recognition at different times. *Neurocomputing*, vol. 212, no. C, pp. 36-47.

Girdhar, R., Ramanan, D. (2017). Attentional Pooling for Action Recognition'' Proc. *Inernational. Conference, Neural Information Processing Systems* , Long Beach, CA, USA, pp. 33–44.

He, T., Mao, H., Yi, Z .(2017) Moving object recognition using multi-view three-dimensional convolutional neural networks. *Neural Computing and Applications*, Springer, vol. 28, no. 12, pp 3827–3835.

Holte, M.B., et al . (2012a). A local 3-d motion descriptor for multi view human action recognition from 4-d spatiotemporal interest points. *IEEE Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 553–565.

Holte, M.B., et al.(2012b). Human pose estimation and activity recognition from multi view videos: Comparative explorations of recent developments. *IEEE journal of selected topics in signal processing*, vol. *6*, no.5, pp.538–552.

Hongyang, L., Jun, C., Ruimin, H.(2017). Multiple Feature Fusion in Convolutional Neural Networks for Action Recognition. *Wuhan University Journal of Natural Sciences*., vol. 22, no. 1, *pp.* 73–78.

Horn, B.K.P., Schunck, B.G.(1981). Determining optical flow. *Elsevier Artificial Intelligence*, vol. 17, no. 1–3, pp. 185-203.

Ijjina, E.P., Chalavadi, K.M.(2016).'Human action recognition using genetic algorithms and convolutional neural networks. *Pattern Recognition*, vol. 59, pp.199–212.

Ji, S., et al. (2013). 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 35, no. 1, pp. 221 – 231.

Karpathy, A., et al. (2014). Large scale video classification with convolutional neural networks. Proc. *Inernational. Conference, IEEE on Computer Vision and Pattern Recognition*, Columbus, Ohio, pp.1725–1732.

Khalajzadeh, H., Mansouri, M., Teshnehlab M. (2014) Face Recognition Using Convolutional Neural Network and Simple Logistic Classifier. Proc. *Inernational. Conference. The 17th Online World Conference on Soft Computing in Industrial Applications Springer*, Cham, vol. 223, pp.197–207.

Klaser, A., Marszalek, M., Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. Proc. *Inernational. Conference, BMVC 19th British Machine Vision*, Leeds, United Kingdom, pp.275:1-10.

Lawrence, S., et al. (1997). Face recognition: a convolutional neural network approach. *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113.

LeCun, Y., et al. (1990). Handwritten digit recognition with a backpropagation network. ,*Advances in Neural Information Processing Systems*, vol. 2, pp. 396–404.

LeCun, Y. (1998). Gradient based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, no. 11, pp.2278–2324.

LeCun, Y., Huang, F.-J., Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. Proc. *Inernational. Conference, IEEE Computer vision and pattern recognition*, Washington, USA, pp. 97-104.

Lim, A., et al. (2017). Real-time optical flow-based video stabilization for unmanned aerial vehicles. *Journal of Real-Time Image Processing*, Springer, pp. 1-11.

Liu, C., et al. (2008). Human-assisted motion annotation. Proc. *Inernational. Conference, IEEE Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, USA, pp. 1-8.

Liu, H., Tu, J., Liu ., M. (2017). Two-Stream 3D Convolutional Neural Network for Human Skeleton-Based Action Recognition. *arXiv preprint: 1705.08106*.

Liu, Z., Zhang, C., Tian, Y. (2016). 3D-based Deep Convolutional Neural Network for action recognition with depth sequences. *Image and Vision Computing*, vol. 55, no. P2 , pp. 93–100.

Matsugu, M., et al. (2003). Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks, Elsevier*, vol. 16, no.5–6, pp. 555–559.

Munaro, M., et al. (2013). 3D flow estimation for human action recognition from colored point clouds. *Biologically Inspired Cognitive Architectures*, vol. 5, pp. 42– 51.

Pablos, A. et al . (2016). Flexible human action recognition in depth Video sequences using masked joint trajectories. *EURASIP Journal on Image and Video Processing*, Springer, vol. 20 , no. 1AB.

Scovanner, P., Ali, S., Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. Proc. *Inernational. Conference*, Multimedia, *ACM*, Augsburg, Germany, pp. 357–360.

Sheikh, Y., Sheikh, M., Shah, M. (2005). Exploring the Space of a Human Action. Proc. *Inernational. Conference, IEEE Computer Vision*, pp. 144-149.

Simonyan, K., .Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Proc. *Inernational. Conference, NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 1, pp. 568–576.

Subramanian, K., Radhakrishnan, V.B., Sundaram, S. (2014). An Optical Flow Feature and McFIS Based Approach for 3-dimensional Human Action Recognition. Proc. *Inernational. Conference, Intelligent Sensors, Sensor Networks and Information Processing IEEE* , pp. 1-6.

Taylor, G., et al. (2010). Convolutional learning of spatio-temporal features'' Proc. *Inernational. Conference. IEEE European Conference on Computer Vision*, Heraklion, Crete, Greece, pp. 140–153.

Tran, D., et al. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. Proc. *Inernational. Conference, IEEE on Computer Vision*, Santiago, Chile, pp. 4489-4497.

Turaga, P. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488.

Willems, G., Tuytelaars, T., Gool, L.V. (2008). An efficient dense and scale invariant spatiotemporal interest point detector. Proc. *Inernational. Conference, European Conference on Computer Vision*, Springer, France, Marseille, pp. 650-663.

Yu, S., et al. (2017). Stratified pooling based deep convolutional neural networks for human action recognition. *Multimedia Tools and Applications*, Springer, vol. 76, no. 11, pp. 13367-13382.