International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:6, 2013

Human Action Recognition Based on Ridgelet Transform and SVM

A. Ouanane and A. Serir

***Abstract*—**In this paper, a novel algorithm based on Ridgelet Transform and support vector machine is proposed for human action recognition. The Ridgelet transform is a directional multi-resolution transform and it is more suitable for describing the human action by performing its directional information to form spatial features vectors. The dynamic transition between the spatial features is carried out using both the Principal Component Analysis and clustering algorithm K-means. First, the Principal Component Analysis is used to reduce the dimensionality of the obtained vectors. Then, the k-means algorithm is then used to perform the obtained vectors to form the spatio-temporal pattern, called set-of-labels, according to given periodicity of human action. Finally, a Support Machine classifier is used to discriminate between the different human actions. Different tests are conducted on popular Datasets, such as Weizmann and KTH. The obtained results show that the proposed method provides more significant accuracy rate and it drives more robustness in very challenging situations such as lighting changes, scaling and dynamic environment.

***Keywords*—**Human action, Ridgelet Transform, PCA, K-means, SVM.

I. Introduction

HUMAN action recognition is one of the most active research areas in computer vision in the last decade, due to its promising applications such as video surveillance, content-based video retrieval, and human-robot interaction.

The action is considered as a set of motion pattern for single period, such as a walking, hand-waving and boxing. The term action and behavior are used interchangeably in order to avoid any confused with the terms in this paper. Broadly, the actions are usually described into two main categories, shape-based and motion-based methods. Shape-based features have been commonly used in behavior and action recognition which are mainly based on the silhouette information [1], [2]. On the other hand, the motion-based methods are based on motion features to characterize human actions [3], [4]. Although the use of such features leads to satisfactory action recognition results, but their computation is expensive.

Many techniques can be used for the human action analysis and recognition. One could represent it by using pixel-level representation or object-level description. The main challenges lie in the extraction of highly descriptive and discriminative features for a minimal compact representation of human action and activity taken into consideration the dynamic environment constraints. In such case, the invariant spatio-temporal descriptors have shown particular promise to characterize the human action due to its rich descriptive power.

Recently, the Radon transform has been used as an invariant spatio-temporal descriptor to characterize the human actions. This transform is defined as summations of image pixels over a certain set of lines. Radon transform provides a mean for determining inner structure of an object. It allows analyzing signal in detail by means of transforming the original signals from the spatial domain into projection space [5]. Among of existing works Singh et al [6] are used Radon transform for pose recognition. But, their work was restricted to hand gestures and feet positions. Wang et al.[7] are then used Radon transform for action recognition. They prove that Radon transform is may be considered as invariant descriptor which the features are invariant to the translation, rotation and scaling. However, the Radon transform has a high-computational cost in term of computing time due to its redundancy.

In this case, we propose a novel algorithm based on Ridgelet transform which aims to outperform the aforementioned transform in terms of invariance and dimensionality. The Ridgelet transform is a directional multi-resolution transform and proves to be very effective in many recent works such as compression [8] and watermarking data [9]. The innovative side of this work consists to use the performance of the Ridgelet transform to represent the spatial information of the human actions by few representative coefficients to overcome the Radon transform redundancy.

The remainder of this paper is organized as follows. The proposed method is presented in Section II. We illustrate experimental results with performances evaluation in Section III. We conclude with a summary and description of future work in Section IV.

II. Proposed Method

In this section, we discuss in detail about the techniques that are used for the human action recognition. The Fig. 1 shows the flowchart of the proposed method. To better understand the proposed algorithm, its main processes are described by the following stages.

The first stage consists of the Background subtraction. It aims to extract the silhouette from video sequence. In this work, we do not focus to deal the background subtraction, which is a widely studied on various literatures.

A.Ouanane is with Laboratory of Image processing and radiation (L.T.I.R.), University of Sciences and Technology Houari Boumediene, BP 32 EL ALIA 16111 Bab Ezzouar Algiers Algeria (e-mail: aouanane@ usthb.dz).

A. Serir is with Laboratory of Image processing and radiation (L.T.I.R.), University of Sciences and Technology Houari Boumediene, BP 32 EL ALIA 16111 Bab Ezzouar Algiers Algeria (e-mail: aserir@ usthb.dz).

Second, we generate the spatial features by using Ridgelet transform.

Third, we perform the temporal relation between the obtained spatial features to form spatio-temporal pattern by using both PCA and K-means algorithms.

Finally, a Support Vector machine classifier is used to classify between the different human actions.

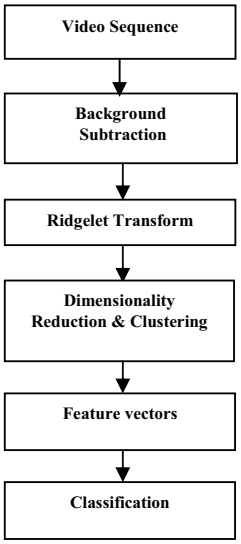

Fig. 1 Flowchart of the proposed Method.

A. Ridgelet Transform

Herein, the Ridgelet transform is introduced by presenting its basic ideas and then we demonstrate how we can generate the spatial features in order to better characterize the human behavior by performing the performance of the proposed approach.

The Ridgelet transform [10], [11] is considered as wavelet analysis in the Radon space. Recall that the Radon transform of an image f is the collection of line integrals indexed by $(\theta,\rho)\in[0,2\pi]\times\Re$ given by:

$$\Re_f(\theta,\rho)=\int f(x,y)\ \ \delta(x\cos\theta+y\sin\theta-\rho)dxdy. \qquad (1)$$

where δ is the Dirac distribution.

Then the Ridgelet transform is precisely the application of a one-dimensional wavelet transform to the slices of the Radon transform where the angular variable θ is constant and ρ is varying.

The Fig. 2 shows an example of Radon transform conducted for 4 different directions (horizontal, vertical, right-diagonal and left-diagonal). One can see that the Radon transforms have the intensive coefficients relative to the orthogonal direction of original images. Furthermore, the horizontal line of original image is represented by the intensive coefficients in radon space with angle equal to $\theta=90°$ in Radon space. Also, it is valid for the vertical, right-diagonal and left diagonal images on which their coefficients are accentuated relative to the angles 0°, 135°, 45° respectively.

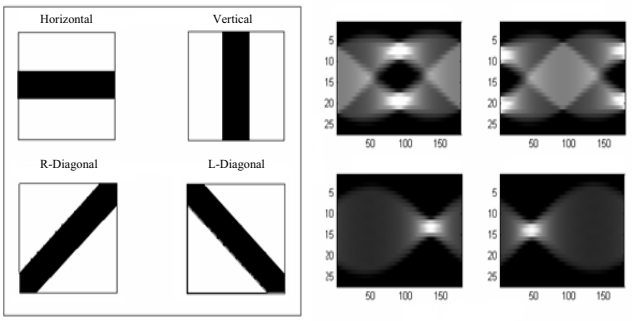

Fig. 2 Radon Transform for horizontal, vertical, right diagonal and left diagonal lines respectively

The Ridgelet transform is carried out by computing one-dimensional wavelet transform along the radial variable in Radon space, $R_f(\theta,.)$, by integrating the variable ρ. Therefore, the Ridgelet transform of image $f\in L^2(\Re^2)$ is described:

$$RT_f(a,b,\theta)=\frac{1}{\sqrt{a}}\int_R\psi\left(\frac{\rho-b}{a}\right)R_f(\theta,\rho)d\rho. \qquad (2)$$

where a is positive and defines the scale and b is any real number and defines the shift, θ is the angle of projection and ψ is the wavelet function. The Ridgelet transform coefficients can be written by the following formula:

$$RT_f(a,b,\theta)=\int_{R^2}\psi_{a,b,\theta}(x,y)f(x,y)dxdy. \qquad (3)$$

with

$$\psi_{a,b,\theta}(x,y)=\frac{1}{\sqrt{a}}\psi\left[\frac{x\cos(\theta)+y\sin(\theta)-b}{a}\right]. \qquad (4)$$

In other words, it corresponds to the projection of the image f on the basis $\psi_{a,b,\theta}$. Hence, the image f is reconstructed by using the following equation:

$$f(x,y)=\int_0^{2\pi}\int_{-\infty}^{\infty}\int_0^{\infty}RT_f(a,b,\theta)\psi_{a,b,\theta}(x,y)\frac{da}{a^3}db\frac{d\theta}{4\pi}. \qquad (5)$$

In this work, we have used a fast implementation of the Radon transform which is performed by the 'Fourier Slice Theorem'.

First, the 2D straight lines equal to the selected number of projections, each line passing through the origin of the 2-D frequency space, which slope equal to the projection angle, and a number of interpolation point equal to the number of rays per projection. The one-dimensional inverse Fourier transform of each interpolated array is then evaluated along each radial followed by a one-Dimensional wavelet transform. The figure 3 shows the flowgraph of the Ridgelet transform.

It is well-known that the Ridgelet transform can provide substantial advantages. First, it is optimal to find lines and segments as well as characterize the movement of specific limbs of the persons in order to model their behaviors based on directional information in Radon space. By performing the resolution-level of the 1-D wavelet transform, the dimensionality of the obtained vectors is slightly reduced by using only the approximation wavelet coefficients. Last and not least, it is also invariant to scaling and translation which provides us invariant spatial features to human behavior analysis. The Fig. 4 and 5 show hand-waving and boxing actions in Radon space respectively.

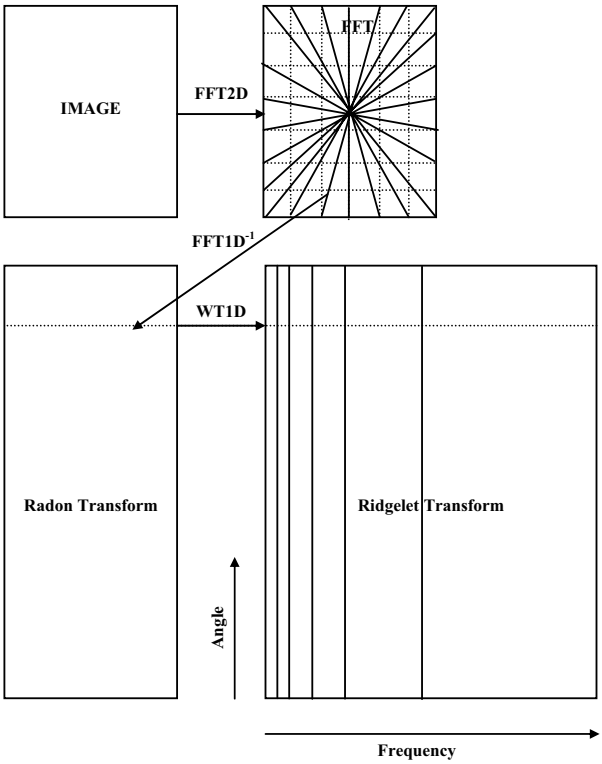

Fig. 3 Ridgelet Transform flowgraph [11]

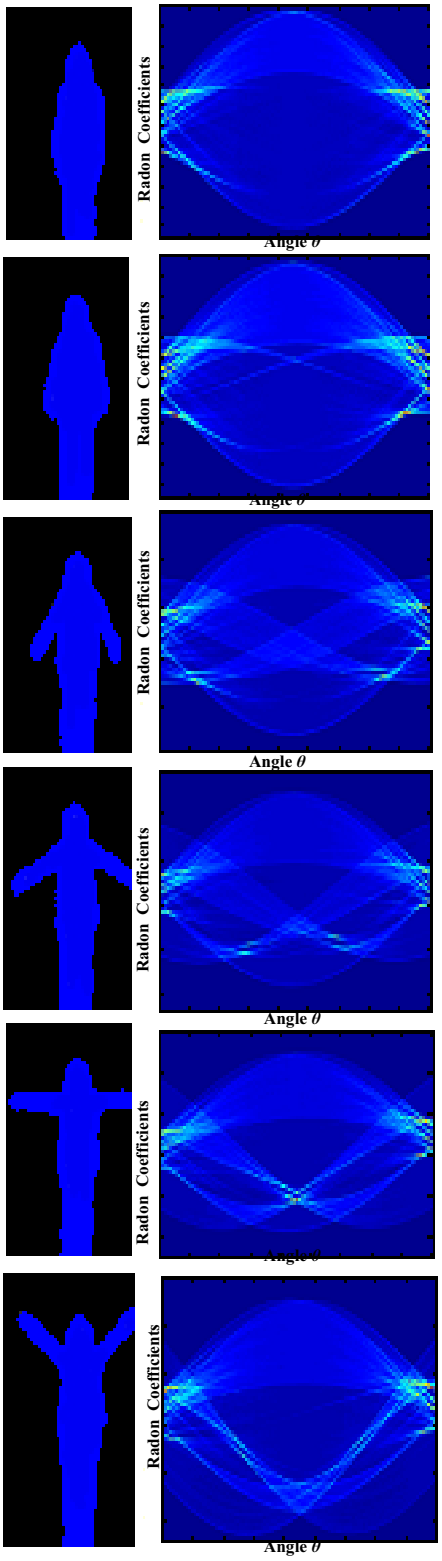

Fig. 4 Hand-waving action in Radon space

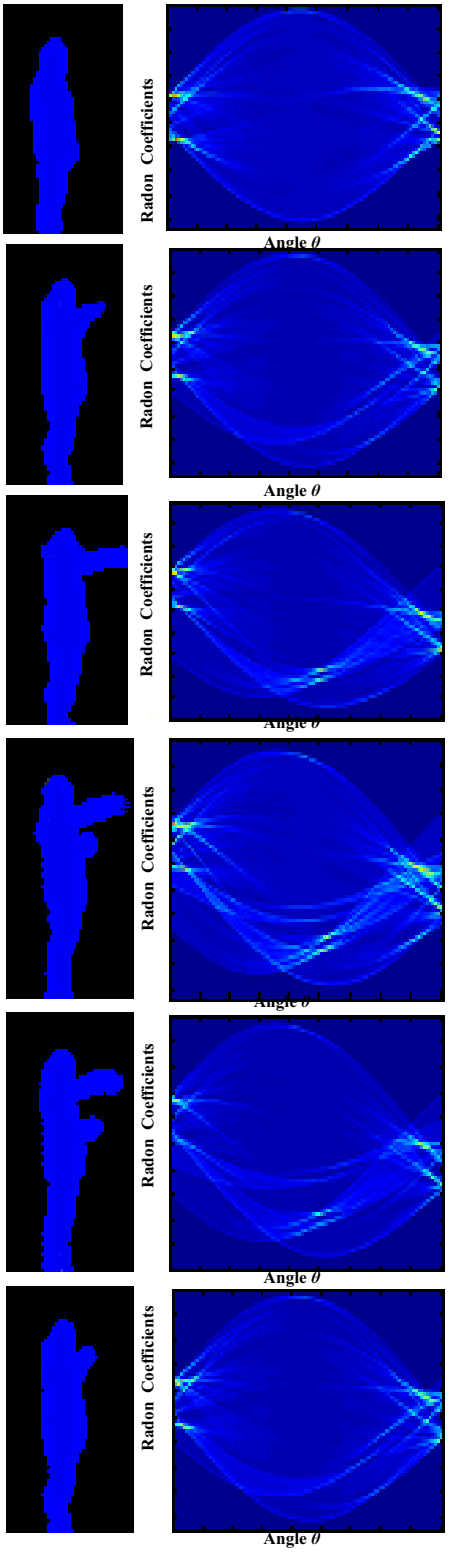

Fig. 5 Boxing action in Radon space

The provided Radon coefficients are performed by one-dimensional wavelet transform with third resolution level to obtain Ridgelet coefficient. The Fig. 6 shows the Ridgelet transform for hand-waving.

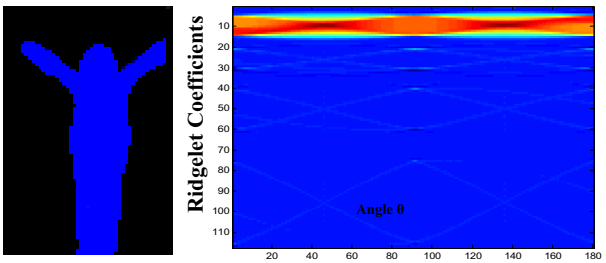

Fig. 6 Hand-waving (left), Ridgelet transform (right).

B. Bag-of-Labels Generation

In this work, the human action is considered as a set of spatial information performed by one frame and it is led by a single person. The dynamic transition is ensured by the temporal relationship between the different spatial information during the evolution of human behavior in time.

The spatial information has been extracted by using the aforementioned descriptor based on Ridgelet transform. Formally, a given action {boxing, hand-clapping, walking,…etc} on video sequence which has N frames is denoted by $f = \{ f_1 ,f_2 ,f_3 , \cdots ,f_i , \cdots | 1 \leq i \leq N \}$. Where, N is relative to the periodicity of action on video sequence and it has been set at N=50 *frames* in this work. The extracted spatial features using Ridgelet transform is denoted by set of m vectors, wherein $R^m = \{ R_1 ,R_2 ,R_3 , \cdots ,R_i , \cdots | 1 \leq i \leq N \}$ and m is the number of training examples in classification stage by SVM. However, the provided vectors R^m are characterized by a high-dimensionality which is may be cause poor recognition rate and high computational cost.

In such case, it is necessary to reduce the dimensionality of the aforementioned vectors R^m on which we have proposed a technique based both the Principal Component Analysis and k-means algorithm. The interest of this technique is a twofold; reduce the dimensionality and characterize the temporal relationship of human behaviors. First, the Principal Component Analysis (PCA) [12], [13] is used to reduce the dimensionality by extracting the pertinent information from the Ridgelet coefficients vectors. The basic concept of this algorithm is conceptually quite simple. The N-dimensional mean vector μ and $d \times d$ covariance matrix Σ are computed for the full data denoted $R^m = \{ R_1 ,R_2 ,R_3 , \cdots ,R_i , \cdots | 1 \leq i \leq N \}$. Then, the eigenvectors and eigenvalues are computed and sorted according to decreasing eigenvalue. After using PCA algorithm, the dimension of R^m including R_i has been properly reduced. The obtained vectors by the PCA algorithm are then performed by using a clustering algorithm K-means [14].

It is well-known that the obtained vectors using PCA are uncorrelated on which the local minima problem of k-means algorithm is no longer raised. The basic idea of this algorithm consists to cluster the provided vectors in the different large of clusters and then encodes each one by the label (index) on which it belongs. These labels are then regrouped to form a set of ordered or unordered labels called set-of-labels. The

latter is considered as a training input in classification stage.

The Fig. 7 shows the different process used to generate the spatio-temporal behavior features.

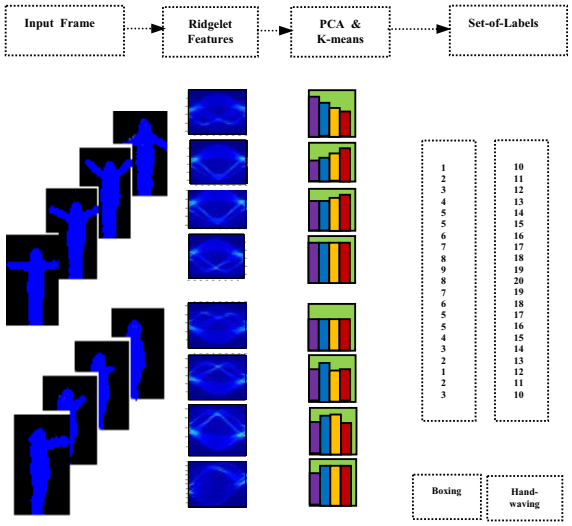

Fig. 7 Spatio-temporal features pattern.

C. Classification Stage

To outperform the decision-making of the proposed algorithm, it is necessary to find an optimal learning algorithm which aims to better discriminate between the different human actions. Among of existing learning algorithms can be used for classification problems is the support vector machine [15]. It has stronger theory interpretation and better generalization performance than the other approaches. Also, it can be considered a linear approach for high-dimensional feature spaces.

Using kernels, all input data are mapped nonlinearly into a high- dimensional feature space. Separating hyperplanes are then constructed with maximum margins, which yield a nonlinear decision boundary in the input space. Using appropriate kernel functions, it is possible to compute the separating hyperplanes without explicitly mapping the feature space. Moreover, it needs lots of labeled data to train classifier model.

Among of existing kernels that can be used in Support Vector Machines models include linear, polynomial, radial basis function (RBF) and sigmoid function. In this work, the RBF kernel has been opted because of their localized and finite responses.

III. Experimental Results

The main goal of this work is to discriminate between the different human behaviors from a dataset. Among of existing popular datasets, we have opted to use Weizmann and KTH datasets. The Weizmann dataset [16] presents a collection of 90 single view video sequences, showing 9 different people, each performing 10 natural actions such as “bend”, “run”, “walk”, “skip”, “jack”, “jump”, “pjump”, “side”, “wave2” and “wave1”. It presents only one scenario, static background, different action classes and inter-class similarity in the local motion, e.g. the jump and skip actions are very similar to each other. The KTH dataset [17] consists of 25 subjects {6 female, 19 male} performing 6 different actions {boxing, hand-clapping, jogging, running, walking, hand-waving. Different scenarios are performed: indoors, outdoors, outdoors with scale variation and outdoors with different clothes.

We evaluate the performance of the proposed by using One-Versus-One approach for the Multi-class SVM model classifier [18], [19] on which we fed each input vector with its related class. The different experiments have been conducted subject to the simple cross-validation technique (leave-one-out strategy) both Weizmann and KTH datasets. All programs have been built by using Matlab 2009 language. The table I shows the confusion matrix of accuracy rate conducted on the Weizmann dataset. We can see that the proposed algorithm is able to discriminate between the different actions with significant accuracy. However, it is noticed that there are slightly confusion between jump, run and skip actions, which is reasonable because they have a similar distribution topic in some scenarios.

Table II illustrates the accuracy rate of six classes {hand-clapping, hand-waving, jogging, running and walking} of the KTH dataset. The reported results on this table prove the effectiveness of the proposed algorithm in terms of robustness to invariant features and decision on which the different classes are well separated. However, a slightly confusion is appeared between the boxing and hand-clapping actions in some scenarios of KTH dataset.

TABLE I
THE CONFUSION MATRIX OF WEIZMANN DATASET

	Bend	Jack	Jump	Pjump	Run	Side	Skip	walk	Wave1	Wave2
Bend	100	0	0	0	0	0	0	0	0	0
Jack	0	100	0	0	0	0	0	0	0	0
Jump	0	0	84	0	0	0	16	0	0	0
Pjump	0	0	0	100	0	0	0	0	0	0
Run	0	0	0	0	92	0	8	0	0	0
Side	0	0	0	0	0	100	0	0	0	0
Skip	0	0	14	0	8	0	78	0	0	0
Walk	0	0	0	0	0	0	0	100	0	0
Wave1	0	0	0	0	0	0	0	0	100	0
Wave2	0	0	0	0	0	0	0	0	0	100

TABLE II
THE CONFUSION MATRIX OF KTH DATASET

	Box	clap	wave	jog	run	walk
boxing	98.6	1.3	0.1	0	0	0
clap	1.4	98.6	0	0	0	0
wave	0.1	0	99.9	0	0	0
jog	0	0	0	100	0	0
run	0	0	0	0	100	0
walk	0	0	0	0	0	100

Furthermore, a comparison between the proposed algorithm and some competitive works on human action recognition is reported in table III. The average accuracy rate of the proposed method is slightly better than the method reported by Liu and Shah [20] with accuracy rate equal to 95.6%. It is also more significant than the methods reported by Laptev [21] and Schindler [22]. The main advantage of using SVM lies in the robustness of decision with a few of training examples unlike other machine learning algorithms which they are required a huge training database wherein they lead a high computational time.

TABLE III
COMPARISON OF RECOGNITION RATE WITH DIFFERENT WORKS ON KTH DATASET

	Accuracy Rate (%)	Classifier type	Year
Proposed method	95.6	SVM	
Liu & Shah [20]	94.2	2-NN	2008
Laptev et al [21]	91.8	SVM	2008
Schindler et al[22]	92.7	SVM	2004

IV. CONCLUSION

In this paper, a novel descriptor based on Ridgelet transform has been presented in order to better characterize the human actions. The spatio-temporal behavior is characterized on two stages; spatial and temporal analysis.

May substantial advantages can be raised by using the proposed algorithm. The spatial features are invariant to translation and scaling due to the performance of the Ridgelet transform. The technique based on PCA approach and clustering algorithm k-means performs significant spatio-temporal features in terms of representativity and dimensionality. The testing stage by using SVM classifier proves the effectiveness of the proposed method in terms of accuracy rate and computational cost. The different experimental results show that the proposed algorithm provides satisfactory performance with significant accuracy rate. For further research, we suggest to use the uniform quantization to quantify the directional information to overcomplete the decision system of the human action recognition.

REFERENCES

[1] J. C. Niebles and F.Li. “A hierarchical model of shape and appearance for human action classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Jun.17–22,pp. 1–8. 2007.

[2] C.Rougier, J.Meunier, A. St-Arnaud, and J.Rousseau. “Fall detection from human shape and motion history using video surveillance,” in Proc.21st Int. Conf. Adv. Inf. Netw. Appl. Workshops, pp. 875–880. 2007.

[3] S. Ali and M. Shah. “Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning”, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Volume 32, Issue 2, pp: 288-303, 2010.

[4] E.B. Ermis, V. Saligrama, P. Jodoin and J. Konrad. “Motion segmentation and abnormal behavior detection via behavior clustering,” in Proc. IEEE Int. Conf. Image Process., Oct. 12–15, pp. 769–772.2008.

[5] J. Li, Q. Pan, H. Zhang, P. Cui. “Image recognition using Radon transform,” Intelligent Transportation Systems,. Proceedings. IEEE, vol.1, no., pp. 741- 744. 2003.

[6] M. Singh, M. Mandal, A. Basu. “Pose recognition using the Radon transform”. In: 48th Midwest Symposium on Circuits and Systems, pp. 1091–1094. 2005.

[7] Y. Wang, K. Huang, T. Tan. “Human activity recognition based on R transform”. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1–8. 2007.

[8] A. Ouanane and A. Serir. “Fingerprint Compression by Ridgelet Transform”, IEEE ISSPIT, 16-19 Décembre 2008 Sarajevo

[9] Z. Zhang, H. Yu, J. Zhang, X. Zhang. “Digital image watermark embedding and blind extracting in the ridgelet domain”, Journal of Communication and Computer, vol.3, No.5.pp.1-7, may 2006.

[10] Candes, E. “ Ridgelets: theory and applications,” Ph.D. thesis, Department of Statistics, Stanford University; 1998.

[11] J. L. Starck, E. J. Candès and D. L. Donoho. The curvelet transform for image denoising. *IEEE Transactions on Image Processing,* vol.11, pp. 670-684. 2000

[12] L. Sirovich and M. Kirby. “Low dimensional procedure for the characterization of human faces”. Journal of the Optical Society of America. A, Optics, Image Science, and Vision, vol 4(3), pp.519–524. 1987.

[13] I.T.Jolliffe. “ Principal component analysis”. Springer, New York. 2002.

[14] Y. Wang, H. Jiang, M. Drew, Z. Li, G.Mori. “Unsupervised discovery of action classes”. IEEE Computer Vision and Pattern Recognition. vol.2, no., pp.1654-1661, 2006.

[15] V.N. Vapnik. “The Nature of Statistical Learning Theory”. New York: Springer-Verlag,1995.

[16] M.Blank, L.Gorelick, E.Shechtman, M.Irani and R. Basri. “Actions as space-time shapes,” in Proc. IEEE Int. Conf. Comput. Vision, pp. 1395–1402. 2005.

[17] C. Schuldt, I. Laptev, and B.I Caputo. “Recognizing human actions: a local SVM approach”. ICPR (17), pp. 32-36, 2004.

[18] C.-W. Hsu and C.-J. Lin. “A comparison of methods for multi-class support vector machines,” IEEE Trans. Neural Netw. , vol. 13, no. 2, pp. 415–425, Mar. 2002.

[19] C.C. Chang and C.J Lin. “LIBSVM: A Library for Support Vector Machines”, 2001.http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[20] J. Liu and M. Shah. “Learning Human Actions via Information Maximization,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

[21] I. Laptev.I, M. Marszał, C.Schmid and B. Rozenfeld. “Learning realistic human actions from movies”. CVPR, pp 1-8, 2008.

[22] K. Schindler and L.V. Gool. “Action snippets: how many frames does human action recognition require”. In: CVPR, pp. 1–8, 2008.

A. Ouanane was born in M’sila, Algeria. He received an engineer degree in electrical engineering from the University of M’sila (Algeria), in 2005. He also received master’s degree in signal and image processing from the University of Sciences and Technology Houari Boumediene (algiers), in 2009. Actually, he is a PhD-candidate and Researcher at the Laboratory of Image processing and radiation (L.T.I.R.), Telecommunications Department, Faculty of Electronics and Computer Science, USTHB University, Algiers, Algeria. He works in computer vision field and their applications including; Human Behavior Analysis, Human Action Recognition, Aggressive Human Behavior Recognition and Tracking of persons.

A. Serir was born in Algiers, Algeria. She received an engineering degree in electrical engineering from state high school, Polytechnic School, Algiers in 1985. She had worked in the design office of Algeria airlines. Then she joined the University of Sciences and Technology and received in 2002, her Ph.D. degree, in image processing. Since, she has been head of the team of ‘‘2D and 3D image processing’’ of the laboratory of image processing and radiation LTIR. She leads several national research projects through integration of bio metrics in smart cards for bank payment or in intelligent video. Her research interests include information processing systems in particular compression and information representation and analysis.