VIDEO BASED HUMAN ACTIVITY RECOGNITION USING WAVELET TRANSFORM AND HIDDEN CONDITIONAL RANDOM FIELDS (HCRF)

Muhammad Hameed Siddiqi[1], La The Vinh[1] and Adil Mehmood Khan[2]
[1]*Ubiquitous Computing Lab, Dept. of Computer Engineering, Kyung Hee University, Suwon, Rep. of Korea*
[2]*Division of Information and Computer Engineering, Ajou University, Suwon, Rep. of Korea*

Keywords: Activity Recognition, Wavelet Transform, HCRF, Video Surveillance.

Abstract: In this research, we proposed testing and validating the accuracy of employing wavelet transform and Hidden Conditional Random Field (HRCF) for video based activity recognition. For feature extraction, Symlet wavelet was tested and decomposed up to 4 levels, and some of the highest coefficients were extracted from each level of decomposition. These coefficients were based on the average frequency of each video frame and the time difference between each frame. Finally, a novel HRCF model was applied for recognition. The proposed method was tested on a database of ten activities, where the data were collected from nine different people, and compared with one of the existing techniques. The overall recognition rate, using the symlet wavelet family (Symlet 4), was 93% that showed an improvement of 13% in performance.

1 INTRODUCTION

Aim of a video-based activity recognition (VAR) system is to automatically recognize a human activity using a sequence of images (video frames). A large number of such systems have been developed in the past, such as (Aggarwal, 1999, Cedras, 1995, Gavrila, 1999, Moeslund, 2006, Turaga, 2008, and Yilmaz, 2006, Siddiqi, 2010).

Feature extraction is an important step in any VAR system. A well-known technique employed for this includes (Uddin, 2008, and 2010). They used Principal Component Analysis (PCA) and Independent Component Analysis (ICA). PCA yields uncorrelated components, especially if the data are a merged of non-Gaussian components then PCA fails to extract components having non-Gaussian distribution (Buciu, 2009). ICA, though better than PCA, is slow to train, especially in the case of high dimensional data. Also, ICA is very weak in managing the inputs.

For recognition, conventional learning methods, such as Hidden Markove Model (HMM), Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Artificial Neural Network (ANN), etc have been mostly employed. Among these, HMM is the most commonly used method (Uddin 2008, 2010). Despite its wide use, it still has some serious deficiencies, such as difficulty to represent multiple interacting activities (Gu, 2009), incapability of capturing long-range or transitive dependencies, and requiring intense training (Kim, 2010).

Our objective was to develop a new feature extraction algorithm and to remove the limitations of the HMM by using a novel hidden condition random field model. Our feature vector is built by extracting the highest wavelet coefficients on the basis of each frame's frequency and the time difference between each frame for each activity. At the recognition stage, the hidden condition random field model is used for acivity recognition.

2 MATERIALS AND METHODS

The aim of this section is to explain the proposed feature extraction and recognition technique.

2.1 Feature Extraction

In this stage, we used the decomposition process applied using Wavelet Transform (WT), for which the video frames were in greyscale. The reason for converting from RGB to gray scale was to improve the efficiency of the proposed algorithm. The wavelet decomposition could be interpreted as

Hameed Siddiqi M., The Vinh L. and Mehmood Khan A..
VIDEO BASED HUMAN ACTIVITY RECOGNITION USING WAVELET TRANSFORM AND HIDDEN CONDITIONAL RANDOM FIELDS (HCRF).
DOI: 10.5220/0003818704010404
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 401-404
ISBN: 978-989-8565-03-7

signal decomposition in a set of independent feature vector. Each vector consists of sub-vectors like

$$V_0^{2D} = V_0^{2D-1}, V_0^{2D-2}, V_0^{2D-3}, \ldots\ldots\ldots, V_0^{2D-n} \quad (1)$$

where V represents the 2D feature vector. If we have a 2D frame x it breaks up into orthogonal sub images corresponding to different visualization. The following equation shows one level of decomposition.

$$X = A_1 + D_1 \quad (2)$$

where X indicates the decomposed image and A_1 and D_1 are called approximation and detail coefficient vectors. If a video frame is decomposed up to multiple levels, the (7) can then be written as

$$X = A_j + D_j + D_{j-1} + D_{j-2} + \ldots + D_2 + D_1 \quad (3)$$

where j represents the level of decomposition, and 'A' and 'D' represent the approximation and detail coefficients respectively. The detail coefficients mostly consist of noise, so for feature extraction only the approximation coefficients are used. In the proposed algorithm, each frame is decomposed up to four levels, i.e., the value of j = 4, because by exceeding the value of j = 4, the image looses significant information, due to that the informative coefficients cannot be detected properly, which may cause misclassification. The detail coefficients further consist of three sub-coefficients, so the (3) can be written as

$$\begin{aligned} X &= A_4 + D_4 + D_3 + D_2 + D_1 \\ &= A_4 + \left[(D_h)_4 + (D_v)_4 + (D_d)_4\right] + \\ &\quad \left[(D_h)_3 + (D_v)_3 + (D_d)_3\right] + \\ &\quad \left[(D_h)_2 + (D_v)_2 + (D_d)_2\right] + \\ &\quad \left[(D_h)_1 + (D_v)_1 + (D_d)_1\right] \end{aligned} \quad (4)$$

Or simply the formula can be written as:

$$X = A_4 + \sum_{i=4}^{1}\left[(D_h)_i + (D_v)_i + (D_d)_i\right] \quad (5)$$

where D_h, D_v and D_d are known as horizontal, vertical and diagonal coefficients respectively. It means that all the coefficients are connected with each other like a chain. Note that at each decomposition step, approximation and detail coefficient vectors are obtained by passing the signal through a low-pass filter and high-pass filter respectively.

After decomposition, the feature vector is created by taking the average of all the frequencies of the activity frames, and also using the time difference between the activity frames. In a specified time window and frequency band with wavelet transform, the frequency is guesstimated. The signal (frame) is analyzed by using the wavelet transform (Turunen, 2011):

$$C(a_i, b_j) = \frac{1}{\sqrt{a_i}} \int_{-\infty}^{\infty} y(t)\psi_{f.e}^{*}\left(\frac{t-b_j}{a_i}\right)dt \quad (6)$$

where a_i is the scale of the wavelet between lower frequency and upper frequency bounds to get high decision for frequency estimation, and b_i is the position of the wavelet from the start and end of the time window with the spacing of signal sampling period. Other parameters include: time t; the wavelet function $\Psi_{f.e.}$, which is used for frequency estimation; and $C(a_i, b_i)$, which are the wavelet coefficients with the specified scale and position parameters. Finally, the scale is converted to the mode frequency, f_m:

$$f_m = \frac{f_a(\psi_{f.e.})}{a_m(\psi_{f.e.}).\Delta} \quad (7)$$

where $f_a(\Psi_{f.e.})$ is the average frequency of the wavelet function, and Δ is the signal sampling period.

2.2 Recognition

To overcome the limitations of HMM and Gaussian mixture HCRF model, we explicitly included a mixture of Gaussian distributions in the feature functions, thus our feature functions could be described in the following forms

$$f_s^{\mathrm{Prior}}(Y, \bar{S}, X) = \delta(s_1 = s)\forall s \quad (8)$$

$$f_{s,s'}^{Transation}(Y, \bar{S}, X) = \sum_{t=1}^{T}\delta(s_{t-1} = s)\delta(s_t = s')\forall s, s' \quad (9)$$

$$f_s^{Observation}(Y, \bar{S}, X) = \sum_{t=1}^{T}\log\left(\sum_{m=1}^{M}\Gamma_{s,m}^{Obs}N(x_t, \mu_{s,m}, \Sigma_{s,m})\right)\delta(s_t = s) \quad (10)$$

$$N(x, \mu_{s,m}, \Sigma_{s,m}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_{s,m}|^{\frac{1}{2}}}\exp\left(-\tfrac{1}{2}(x-\mu_{s,m})'\Sigma_{s,m}^{-1}(x-\mu_{s,m})\right) \quad (11)$$

where M is the number of density function, D is the dimension of the observation, $\Gamma_{s,m}^{Obs}$ is the mixing weight of the m^{th} component with mean $\mu_{s,m}$ and covariance matrix $\sum_{s,m}$. As we can see in (15), Γ, μ, and $\sum$ we can be updated during the training phase, hence we can set

$$\Lambda_{s,m}^{Obs} = 1 \forall s \tag{12}$$

As a result, the conditional probability can be rewritten as below

$$p(Y \mid X;\Lambda,\Gamma,\mu,\Sigma) = \frac{\sum_{\bar{S}} \exp\left(\begin{array}{l} \sum_{s} \Lambda_{s}^{Prior} f_{s}^{Prior}(Y,\bar{S},X) + \\ \sum_{s s'} \Lambda_{s s'}^{Transition} f_{s s'}^{Transition}(Y,\bar{S},X) + \\ \sum_{s} f_{s}^{Observation}(Y,\bar{S},X) \end{array} \right)}{z(X;\Lambda,\Gamma,\mu,\Sigma)} \tag{13}$$

$$= \frac{\sum_{\bar{S}=S_1,S_2,\ldots,S_T} \exp\left(\Lambda_{S_1}^{Prior} + \sum_{t=1}^{T} \left(\Lambda_{S_{t-1},S_t}^{Transition} + \log\left(\sum_{m=1}^{M} \Gamma_{S_t,m}^{Obs} N\left(x_t,\mu_{S_t,m},\Sigma_{S_t,m}\right) \right) \right) \right)}{z(X;\Lambda,\Gamma,\mu,\Sigma)} \tag{14}$$

$$= \frac{Score(Y|X;\Lambda,\Gamma,\mu,\Sigma)}{z(X;\Lambda,\Gamma,\mu,\Sigma)} \tag{15}$$

Based on (19) and (20), we can compute the conditional probability by using the well-known forward and backward algorithm as below

$$\alpha_\tau(s) = \sum_{\bar{S}=S_1,S_2,\ldots,\{S_\tau=S\}} \exp\left(\Lambda_{S_1}^{Prior} + \sum_{t=1}^{T} \left(\Lambda_{S_{t-1},S_t}^{Transition} + \log\left(\sum_{m=1}^{M} \Gamma_{S_t,m}^{Obs} N\left(x_t,\mu_{S_t,m},\Sigma_{S_t,m}\right) \right) \right) \right)$$

$$= \sum_{s'} \alpha_{\tau-1}(s') \exp\left(\Lambda_{s's}^{Transition} \log\left(\sum_{m=1}^{M} \Gamma_{s,m}^{Obs} N\left(x_\tau,\mu_{s,m},\Sigma_{s,m}\right) \right) \right) \tag{16}$$

$$\beta_\tau(s) = \sum_{\bar{S}=\{S_\tau=S\},S_{\tau+1},\ldots,S_T} \exp\left(\Lambda_{S_1}^{Prior} + \sum_{t=1}^{T} \left(\Lambda_{S_{t-1},S_t}^{Transition} + \log\left(\sum_{m=1}^{M} \Gamma_{S_t,m}^{Obs} N\left(x_t,\mu_{S_t,m},\Sigma_{S_t,m}\right) \right) \right) \right)$$

$$= \sum_{s'} \beta_{\tau+1}(s') \exp\left(\Lambda_{s's}^{Transition} \log\left(\sum_{m=1}^{M} \Gamma_{s,m}^{Obs} N\left(x_\tau,\mu_{s,m},\Sigma_{s,m}\right) \right) \right) \tag{17}$$

$$Score(Y \mid X;\Lambda,\Gamma,\mu,\Sigma) = \sum_{S} \alpha_T(s) = \sum_{S} \beta_1(s) \tag{18}$$

In the training phase, our goal was to find the parameters (Λ, Γ, μ, and ∑) to maximize the conditional probability of the training data.

3 RESULTS AND DISCUSSION

In order to evaluate the proposed algorithm, we used a publicly available dataset (Gorelick, 2007), containing ten activities. Each activity is performed by nine different people. The frame size is 144 x 180. The confusion matrix, shown in Table 1, shows the recognition rate of the proposed algorithm.

3.1 Comparison with Existing Algorithm

The proposed method was compared with one of the existing algorithms (Uddin, 2008), whose results are shown in Table 2, in terms of accuracy. The proposed algorithm improved the recognition rate by about 13%. This improvement in accuracy could be attributed to the use of symlet wavelet, which is a compactly supported wavelet with the least asymmetry and the highest number of vanishing moments for a given support width, and HCRF, which provides a better performance because it solves the limitations of the Conditional Random Fields (CRF) HMM.

Table 1: Recognition rates for the testing dataset obtained by the proposed algorithm model (blank cells represent 0%). The average accuracy is 93.0 %.

	bend	jack	pjump	jump	run	side	skip	Walk	wave1	wave2	unknown
bend	**93%**										7%
jack		**94%**								2%	4%
pjump			**92%**			8%					
jump				**86%**	5%		9%				
run				3%	**92%**		5%				
side			6%	5%		**89%**					
skip							**93%**				7%
walk					6%			**94%**			
wave1									**100%**		
wave2		3%								**97%**	

Table 2: Recognition rates for the testing dataset obtained by the existing algorithm (Uddin, 2008). The blank cells represent 0%. The average accuracy is 80.0 %.

	bend	jack	pjump	jump	run	side	skip	walk	wave1	wave2	unknown
bend	**83%**										17%
jack		**82%**								12%	6%
pjump			**80%**			18%					2%
jump				**78%**	11%		11%				
run				7%	**74%**		12%	7%			
side			15%	10%		**75%**					
skip				3%			**79%**				18%
walk					20%			**80%**			
wave1		6%							**84%**		10%
wave2		15%								**85%**	

4 CONCLUSIONS

In this research, we proposed a VAR algorithm that employs wavelet transform and HCRF model. The proposed algorithm was tested on a publicly available dataset. The recognition results were compared with one of the existing techniques that used PCA, ICA, and HMM. The overall recognition rate using the symlet wavelet family (Symlet 4) was 93%. These results showed an improvement of 13% in performance.

ACKNOWLEDGEMENTS

This work was supported by the new faculty research fund of Ajou University.

REFERENCES

Aggarwal, J. K., Cai, Q., 1999. Human motion analysis: A review. *Comput. Vis. Image Und*., vol. 73(3), pp. 428-40.

Cedras, C., Shah, M., 1995. Motion-based recognition: A survey. *Image Vis. Comput*., vol. 13(2), pp. 129-55.

Gavrila, D. M., 1999. The visual analysis of human movement: a survey: *Comput. Vis. Image Und*., vol. 73(1), pp. 82-98.

Gorelick, L., Blank M., Shechtman E., Irani M., Basri R., 2007. Actions as Space-Time Shapes. *IEEE Trans. PAMI*., vol. 29(12), pp.2247-53.

Gu, T., Wu, Z., Tao, X., Pung, H. K., Lu, J., 2009. epSICAR: An Emerging Patterns based approach to sequential, interleaved and Concurrent Activity Recognition. *In Proc. of IEEE Intl. Conference on Pervasive Computing and Communications*.

Kim, T.-S., Uddin, M. Z., 2010. Silhouette-based Human Activity Recognition Using Independent Component Analysis, Linear Discriminant Analysis and Hidden Markov Model. *New Developments in Biomedical Engineering*, ISBN: 978-953-7619-57-2. Edited by: Domenico Campolo. Published by InTech.

Moeslund, T. B., Hilton, A., Kruger, V., 2006. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Und*., vol. 104(2), pp. 90-126.

Siddiqi, M. H., Fahim, M., Lee, S. Y., Lee, Y.-K, 2010. Human Activity Recognition Based on Morphological Dilation followed by Watershed Transformation Method. *Proc. of International Conference on Electronics and Information Engineering (ICEIE)*, pp. V2 433-V2 437.

Turaga, P., Chellappa, R., Subrahmanian, V. S., Udrea, O., 2008. Machine Recognition of Human Activities: A survey. *IEEE Trans. Circuits and Systems for VideoTechnology*, vol. 18(11), pp. 1473-88.

Turunen, J, 2011. A Wavelet-based Method for Estimating Damping in Power Systems. PhD. Thesis, Aalto University, School of Electrical Engineering, Department of Electrical Engineering Power Transmission Systems.

Uddin, M. Z., Lee, J. J., Kim, T.-S, 2010. Independent shape component-based human activity recognition via Hidden Markov Model. *Appl. Intell*, vol. 33(2), pp. 193-206.

Uddin, M. Z., Lee, J. J., Kim, T.-S. 2008. Shape-Based Human Activity Recognition Using Independent Component Analysis and Hidden Markov Model. *Proc. of 2[1st] International Conference on Industrial, Engineering, and other Applications of Applied Intelligent Systems*, pp.245-254.

Yilmaz, A., Javed, O., Shah, M., 2006. Object tracking: A survey. *ACM Comput. Surv*., vol. 38(4).