# ACTIVITY MODELING WITH SPATIO-TEMPORAL TEXTURE PRIMITIVES

*Burcu Kepenekci, Gozde B. Akar*

Middle East Technical University,
Department of Electrical and Electronics Engineering, Ankara, Turkey

## ABSTRACT

In this paper, a novel structure is proposed for human activity modeling using time sequential spatio-temporal texture primitives. Gabor filters, which are proven to be robust 2D texture representation tools, are extended to 3D domain to capture spatio-temporal texture features. A well known filtering algorithm and an unsupervised clustering algorithm, the Genetic Chromodynamics, are combined to select salient spatio-temporal features. Each state of activity is represented as action units with its salient spatio-temporal feature set, which are also the symbols of our codebook. To overcome temporal variation between different performances of the same action, a profile Hidden Markov Model is applied with Viterbi Path Counting (ensemble training). Not only the parameters and the structure but also the codebook is learned during training. Proposed structure enables to recognize sub-sequences and does not require end point constrains. This also provides robustness to the missing data and occlusions.

## 1. INTRODUCTION

Understanding human activities is one of the key issues for the new generation security systems. Detection and tracking systems have progressed significantly in the past few years, making human motion and behavior interpretation, naturally, the following step.

Despite the fact that good results were achieved by state of the art activity recognition approaches, these still have some limitations. Early body-part based methods suffer from the computational complexity and the lack of reliable segmentation of body parts in low resolution surveillance videos. Some of the recent approaches have shown that it is useful to analyze activities by looking at a video sequence as a space-time volume (of intensities, gradients, optical flow, or other local features) [1]. Those holistic methods process each frame (or segmented part of each frame as the moving subject) without focusing on the individual body parts and achieve better results on unconstrained environments.

On the other hand some methods also supports holism on time domain by modeling each action as a "bag of features" where others have frame based time sequential representation or state sequential representation based on time segmented video. "Bag of features" representation has the main drawback that different time sequential data can produce the same feature set. On the other hand frame by frame representations suffer from redundancy. State sequential representation is more appropriate for the action recognition problem. However, segmenting video into states and getting a good representation of those states are the two main problems one must solve to achieve a good performance. Frame sequential and state sequential representations must also handle unaligned video sequences, while this is not an issue in "bag of features" representation.

To handle high dimensionality of the video data, most methods apply a simple feature selection based on filtering and/or dimensionality reduction (PCA) algorithm. Other methods employ clustering (SVM, kNN are the two common algorithms), histogramming or sparse sampling of data [2].

Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) are the two common methods used by the approaches based on sequential representation of the activity [1]. DTW deals with sequence aligning by operations of deleting and inserting compression expansion, and substitution, of subsequences. However, it requires rigidly fixed end-points and lacks the consideration of inter actions between nearby subsequences occurring in time. HMM considers this correlation between adjacent time instances by formulating a Markov process. Mapping frames into symbols is the main challenge of HMM approaches. Moreover, construction of a successful HMM structure requires manual design by experts.

Our approach in this paper is based on modeling the activity as a pattern composed of time sequential spatio-temporal texture primitives. Therefore, our first objective is to analyze the texture of the motion and this is achieved by 3D Gabor filtering followed by a two step feature selection algorithm to eliminate redundancy and extract relevant features. 3D Gabor filters are inspired from the human visual system; they give high responses at spatio-temporal locations with high variation at different scales and orientation.

Proposed feature selection algorithm locates the spatio-temporal texture primitives in both spatial and temporal domains. While modeling the activity with these spatio-temporal texture features, their temporal ordering is also used as additional discriminative information between different classes of activities. We adopt a profile HMM structure which has inserting and deleting states to align

different length sequence. Moreover our proposed structure enables to recognize sub-sequences and does not require end point constrains.

Rest of the paper is organized as follows. In Section 2, the proposed spatio-temporal representation is explained, followed by the profile HMM structure. Results are provided in Section 4 and finally we conclude in Section 5.

## 2. SPATIO-TEMPORAL REPRESENTATION

In this work, as a preprocessing step, regions of probe activity are segmented as the foreground, and in the following sections only these regions are processed. We used an adaptive background mixture model for foreground segmentation [7].

### 2.1. 3D Gabor Filtering

3D Gabor filters have very similar aspects with 2D Gabor filters. 3D Gabor filters allow the description of temporal frequency structure and temporal relations, in addition to the description of the spatial frequency structure and spatial relations [3].

A 3D Gabor is the product of a 3D Gaussian and a 3D harmonic function. The length of the axes is controlled by the Gaussian and the frequency is controlled by the harmonic function.

$$G_j(\vec{x}) = Kexp(-\pi\sigma^2\vec{x}^2)\left(\exp\left(j2\pi\vec{k}_j\vec{x}\right) - exp\left(-\frac{\sigma^2}{2}\right)\right) \quad (1)$$

$$\vec{k}_j = (k_v \sin\theta\cos\beta, k_v \sin\theta\sin\beta, k_v \cos\theta) \quad (2)$$

where K scales the magnitude of the Gaussian envelope, $\vec{x}$=(x,y,t) is the location of the peak of the Gaussian envelope, $\vec{k}_j$ spatial and temporal frequencies of the sinusoid carrier in spherical coordinates with magnitude $k_v$ and directions $\beta, \theta$. Expression exp(-$\sigma^2$/2) removes the bias of the filter. By eliminating the DC response filters will not respond to the absolute intensity of the image. 3D Gabor responses are computed by convolving the video sequence and the 3D Gabor kernels.

### 2.2. Feature Selection

In this work a two step feature selection algorithm is applied to find the representative features of the activity: in the first step interest points are found by filtering 3D Gabor responses, in the second step by applying clustering, discriminative spatio-temporal feature clusters are found as the action units of the underlying activity. Locality of the features is assured in both spatial and temporal domains.

Since the proposed feature selection algorithm is fully automatic, numbers of final feature clusters and number of feature vectors of the final feature clusters are not set a priori. Proposed feature selection algorithm generates a representation for each class according to its specific temporal texture properties effectively.

#### *2.2.1. Filtering*

Filtering is applied to 3D Gabor responses to obtain a local spatio-temporal feature map of each frame. 3D Gabor filters give high responses at the locations of sharp changes in curvature in spatio-temporal domain. Since such locations correspond to some low level salient features, by filtering, local peaks of 3D Gabor responses are selected as spatio-temporal interest points.

Feature vectors, v, of each frame in the sequence are generated at the feature points as a composition of Gabor filter response coefficients.

#### *2.2.2. Genetic Chromodynamics Clustering*

Genetic Chromodynamics (GC) is a recent evolutionary unsupervised clustering procedure that maintains population diversity and forces the formation of stable sub-populations [4].

In GC algorithm, each individual participates in the forming of the new generation. Its mate is searched by applying a local selection scheme. If a second chromosome is found within its local range (called the mating region), they recombine and the competition for survival of the fittest is held between the resulting offspring and the first parent only. If no chromosome can be found in the local range of the considered chromosome, it is mutated. The removal of less fit solutions is performed by introducing a special operator that merges very similar chromosomes into one chromosome.

Proposed clustering method utilizes Genetic Chromodynamics procedure for dimensionality reduction by both feature selection and data partitioning.

In this work filtered feature set *(F)* of each frame is considered as an individual. Locality is maintained in both time and space. Similarity between two individuals is computed as the mean similarity of its features.

$$S(v_1,v_2)=\sqrt{\frac{\Sigma|v_1||v_2|}{\Sigma|v_1|^2\,\Sigma|v_2|^2}}\left[abs(d_1-d_2)<\tau_d\right] \quad (3)$$

$$Sim(F_1,F_2)= avg_{\forall\, v_1\in F_1, v_2\in F_2}\, S(v_1,v_2) \quad (4)$$

Each spatio-temporal feature set, *F*, is composed of spatio-temporal feature vectors, *v*, their normalized distance from the center of the foreground region, *d*. $\tau_d$ is the distance threshold.

A weight, *w*, assigned to each vector indicating its appearance frequency, and is set to *N* at the beginning and decreased by 1 at each similarity calculation if there is no match for that feature vector.

Mating region, *MT*, is defined by temporal distance and feature similarity. If *MT($F_t$)* is not an empty set, then the

member, $F_{t+\varepsilon}$, with the lowest similarity is selected as the mate. Selection of the least similar mate avoids local overfitness. Once a mate is selected for $F_t$ crossover is applied to generate an offspring. A match for each feature vector of $F_t$ is search in $F_{t+\varepsilon}$ and the one with the higher response magnitude is selected for the offspring $(\acute{F}_t)$. Then $\acute{F}_t$ and $F_t$ is evaluated; the fittest one is to survive for the next generation. The following evaluating function is used to find the fitness.

$$\text{Eval}_{MT}(\hat{F}) = e^{-\sum_{F_j \notin MT(F_t)} dist(\hat{f},F(j)) + \sum_{F_j \in MT(F_t)} dist(\hat{f},F(j))} \quad (5)$$

Where $dist(F_1, F_2) = 1 - Sim(F_1,F_2)$. This evaluation function supports increased similarity in the mating region and increased distance to the rest of the population.

If $MR(F_t)$ is an empty set, then $F_t$ is mutated. To mutate, weight of the less frequent feature vectors of an individual $F_t$ is set to 0 which indicates that they are eliminated. If weight of a feature is less than a threshold it is considered as noise. Since features are computed in a 3D space *(x,y,t)* if a feature is not locally consistent between consecutive frames it can be stated that it is not related to the probe motion.

By mutation some individuals can be eliminated completely. However our motive is not to assign each feature set $F_t$ to a cluster but to find feature clusters that will effectively represent the underlying motion texture.

Above procedure is applied for every individual in the current population. Before iterating for the next generation, merging is applied. Merging procedure is introduced by GC algorithm to remove less fit individuals, (thus) to decrease the population size. It must be noted that at the mating region only the first parent and the offspring is compete for the survivor. Without merging, at every generation, while the number of similar individuals increases as the result of crossover, total number of individuals would stay the same.

Merging region of $F_t$ is defined similarly as the mating region, but including itself. To combine individuals in the merging region, first all but the best individual from the selection is removed, and then its feature vectors are compared to every feature vector in the merging region; the one that has the higher response magnitude is selected if two feature vectors match. The best individual is selected according to the same evaluation function of survival (5) but instead of mating region, merging region is considered.

At the end of the merging a new generation is obtained. If population does not alter from the previous generation by mating and/or merging procedure is finalized, otherwise mating and merging applied for the new generation as well.

## 3. ACTIVITY RECOGNITION WITH PROFILE HMM

### 3.1. Profile HMM

A Profile HMM is a standard topology for modeling sequence motifs [5]. There are three kinds of states in profile HMM (figure 2): *matching states (M), deletion states (D)* and *insertion states (I).* Matching states are the ones that emit codebook symbols. When there is a region in the model that is not present in the sequence, the model transits to the Deletion state. *Insertion states (I)* handle the cases when the sequence contains a region that is not present in the model. To update the structure, profile HMM introduces model surgery strategy to remove rarely visited matching states and/or insert frequently visited insertion states matching states.

### 3.2. Adapting Profile HMM Procedure for Action Modeling

As the body parts moves, 3D Gabor filters match at some local points during the activity; at a certain instance they produce maximum response showing the points of maximum spatio-temporal variations which characterize the motion texture. Each of such spatio-temporal features is found and clustered into local action units by proposed feature selection algorithm. Temporally ordered feature clusters correspond to the observation sequence. Codebook is obtained as the unique feature clusters. Rather than a standard profile HMM, a profile HMM with a self-looping model is proposed (Fig. 1) to handle sequences with different start and end points by subsequence matching. Proposed profile HMM structure allows local alignment.

Sequence alignment is achieved by applying Viterbi Algorithm to find the optimal state path with current model parameters.

Initially the number of *matching states* is set to the number of clusters in the first sample. When a matching state removed from the model as the result of model surgery, if there exists any symbol exclusive to that matching state it is removed from codebook as well. Self transition probability of the flanking state is set close to 1.

For parameter learning training sequences iteratively inserted into the model structure using Viterbi algorithm to select the path to be incremented. If an unknown symbol is observed it is added to the codebook. Observation symbols are compared to codebook symbols according to the average similarity of their vectors (3). Viterbi Path Counting method is applied, which simply counts states and transitions and estimates the model parameters from those total counts. The dependence upon the initial value is minimized with this method. Model surgery is applied at the end of each iteration.

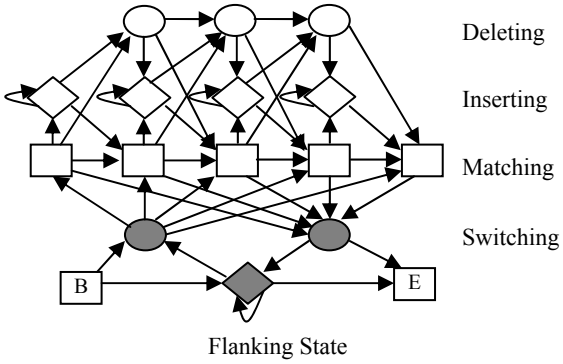



**Fig 1:** Proposed profile HMM structure

Activity recognition with profile HMM is determining whether a new sequence contains the motif. Given the learned model probability of the observed sequence is computed. Although this gives a score, it is also required to analyze the states that generate the score; underlying motif of a human action is captured by matching states. If no symbols were emitted by matching states, then the motif is not present in the observation sequence. The most likely path is found by using the Viterbi back-tracing algorithm.

Each human action class is modeled by a distinct HMM. To recognize an action sample its final scores are calculated for all possible models, and the action whose final score is highest is selected as label.

## 4. RESULTS

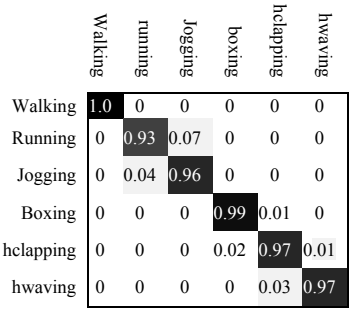
We evaluate the proposed method with two human action databases, namely Weizmann human action database [2] and KTH human action database [6].

The KTH actions dataset consists of six human action classes performed several times by 25 subjects in different scenarios. Each sequence is further divided into a total of 2391 sequences. We use the original training and test split so our results are directly comparable to the recent survey by Wang et al. [2]. Our confusion matrix on KTH database is given in Fig.2.a. Proposed algorithm achieved %96.98 average accuracy on KTH database. Results of other algorithms reported in literature are also given in Fig.2

Weizmann contains two databases: classification database and robustness database. Classification database contains 93 low-resolution 180 x 144 video sequences from nine people, each performing 10 actions. Robustness dataset includes 10 different altered "walk" activity. Our method yields perfect recognition rate that is 100%, on both datasets.

## 5. CONCLUSIONS

In this work a novel spatio-temporal activity representation based on 3D Gabor filters is proposed. Feature selection algorithm locates the spatio-temporal texture primitives in both spatial and temporal domains.




(a)

| Method | Accuracy |
|---|---|
| Our method | 96.98% |
| [8] | 95.8% |
| [9] | 95% |
| [10] | 94.16% |
| [11] | 92.7% |
| [12] | 92.51% |

(b)

**Fig 2:**(a)Confusion matrix, (b) Results reported in literature.

The proposed structure enables to recognize sub-sequences and does not require end point constrains. This also provides robustness to the missing data and occlusions. The proposed method is not domain specific and can be applied to a wide range of activities, and can also be applied to concatenated activities. Experimental results show that the proposed algorithm achieved the best result on KHT dataset, among the methods in the literature.

## 6. REFERENCES

[1] Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O., "Machine Recognition of Human Activities: A Survey", *CirSysVideo(18)*, No. 11, November 2008, pp. 1473-1488.

[2] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009, pp. 127–138.

[3] O. Chomat, J.L. Crowley, "Probabilistic Recognition of Activity using Local Appearance", In *CVPR 1999*, USA.

[4] D. Dumitrescu, "Genetic chromodynamics," Studia Universitatis Babes-Bolyai Cluj-Napoca, *Ser. Informatica*, vol. 45, no. 1, 2000, pp. 39–50.

[5] R. Durbin, S. Eddy, A. Krogh and G. Mitchison. "Biological Sequence Analysis", Cambridge University Press, 1998.

[6] L. Gorelick, et. al, "Actions as space–time shapes", IEEE *Trans. on PAMI, 2009,* vol. 29, pp. 2247–2253.

[7] Bailo, G.; et. al., "Background estimation with Gaussian distribution for image segmentation, a fast approach", *in Proc. Measur. Sys. for Homeland Security, Contraband Detect. and Personal Safety Workshop,* March 2005, pp.2-5

[8] Chen, Ming-Yu and Hauptmann, Alexander, "MoSIFT: Recognizing Human Actions in Surveillance Videos" (2009). Computer Science Department. Paper 929.

[9] Kim, T.K., Wong, S.F., Cipolla, R. "Tensor Canonical Correlation Analysis for Action Classification." IEEE CVPR, 2007: 1-8.

[10] Chang Liu, Pong C. Yuen. "Human action recognition using boosted EigenActions." Image and Vision Computing 28 , 2010: 825–835 .

[11] Schindler, K. and Van Gool, L. "Action snippets:How many frames does human action recognition." IEEE CVPR, 2008.

[12] Mori, Yang Wang and Greg. "Max-Margin Hidden Conditional Random Fields for Human Action Recognition." CVPR, 2009.