Human Action Recognition By Sequence of Movelet Codewords

Xiaolin Feng[†] Pietro Perona [†‡]

† California Institute of Technology, 136-93, Pasadena, CA 91125, USA
‡ Università di Padova, Italy
{xlfeng,perona}@vision.caltech.edu

Abstract

An algorithm for the recognition of human actions in image sequences is presented. The algorithm consists of 3 stages: background subtraction, body pose classification, and action recognition. A pose is represented in space-time – we call it 'movelet'. A movelet is a collection of the shape, motion and occlusion of image patches corresponding to the main parts of the body. The (infinite) set of all possible movelets is quantized into codewords obtained by vector quantization. For every pair of frames each codeword is assigned a probability. Recognition is performed by simultaneously estimating the most likely sequence of codewords and the action that took place in a sequence. This is done using hidden Markov models. Experiments on recognition of 3 periodic human actions, each under 3 different viewpoints, and 8 nonperiodic human actions are presented. Training and testing are performed on different subjects with encouraging results. The influence of the number of codewords on algorithm performance is studied.

1 Introduction

Detecting and recognizing human actions and activities is one of the most important applications in machine vision. There are many approaches recently to address it from different point of view [5, 4, 6, 8, 9, 10, 1, 2, 11, 7, 3]. Both top-down methods starting with human body silhouettes [5] and bottom-up methods based on low level image features such as points [9] and patches [6, 4] have been proposed for detecting the human body and measuring body pose. In the case of exploring patch features, typical bottom-up method assumes the color and/or texture of each body part is known in advance. We do not wish to make any assumptions on the style of clothing (or absence thereof) and/or on the presence of good boundaries between limbs and torso. Therefore we take top-down approach.

Our basic assumption is that a sequence of regions of interest (ROI) containing a foreground moving object is obtained. We discretize the space of poses into a number of codewords which are fitted to ROI for the recognition. In order to further improve the signal-to-noise ratio we consider poses in space-time which are called *movelets*. We model each action as a stochastic concatenation of an arbitrary number of movelet codewords, each one of which may generate images (also stochastically). Hidden Markov models are used to represent actions and the images they generate. The Viterbi algorithm is used to find the most likely sequence of codewords and action that are associated to a given observed sequence. The movelet codewords are learned from training examples and are shared amongst all actions (similarly to an alpabed of letters which are shared by all words).

Results on training of models and recognition of 3 periodic actions and 8 nonperiodic actions imaged from multiple viewpoints are presented. We also explore experimentally what is the necessary number of codewords to represent the movelet space and how action recognition performance varies as a function of the number of frames that are available.

2 Movelet and Codeword: Human Body Modeling

If we only monitor the main body, a human body configuration can be characterized by the shape and motion of the 10 parts: head, torso, 4 parts of upper limb and 4 of lower limb. We name such configuration representation *movelets*. The shape of each corresponding body part on the image plane is modeled as a rectangle S_j (5 degree of freedom (DOF)). Assume this shape deforms to be S_j' (5 DOF) in the next frame. This shape deformation is no longer 3 dimensional rigid motion on the image plane due to loose clothes and occlusion etc., but a non-rigid motion applying in its full parametric space as $S_j^{'} - S_j$. Therefore, a human body movelet can be represented as $\mathbf{M} = (S, S') = (\bigcup_{j=1}^{10} S_j, \bigcup_{j=1}^{10} S_j')$. The advantage of this parameterization over the direct shape and motion representation is its parameters are in the same metric space.

Consider a set of human actions we are interested in. The actions are sequences of movelets. Different actions may contain similar movelets at certain time. We collect the set of the movelets of all actions that have been observed over a learning time period, and cluster them into C codewords denoted as $\mathbf{C}_i, i = 1 \dots N$ using an unsupervised vector quantization method. We choose K-means algorithm in this paper. The vector quantization coarsely divides the whole movelet space into N regions and represents each region by a codeword.

There are two practical issues:

First, Origin Selection: The body movelet is defined relative to an origin which we set at the center of the rectangular shape of the head. The positions of all the parts are relative positions to this origin. The DOF of the movelet is reduced by 2 correspondingly.

The second issue is Self-Occlusion: we observe in our experiments that over half of the movelets have some of the parts occluded. Therefore their dimension is reduced to a lower DOF. We divide the movelets into subgroups according to their occlusion patterns. The vector quantization is performed in each subgroup with the number of codewords proportional to the relative frequency of the occlusion pattern. The union of the trained codewords from subgroups gives the final dataset of codewords.

3 Recognition of A Configuration

To recognize the observed action, we assume that the images of the moving body are segmented as the foreground in image sequence, which sometimes can be done by applying either background subtraction or independent moving object segmentation.

For each pair of frames, an observed configuration is defined as the foreground pixels $\mathbf{X} = \{X, X'\}$, where X and X' respectively represent the 2D positions of the foreground pixels in the first and second image frame. Although we model the human body by parts in the previous section, for the recognition, due to the lack of technique to segment human body parts robustly, the observed configuration consists of foreground pixels in their entirety and top-down approach is applied. Given a codeword with shape S and deformed shape S', we wish to calculate the likelihood of observing X and X'. We assume, given the shape S, apriori probability to detect a pixel X_i in the image as a foreground(*fg*) or background (*bg*) data is a Binomial distribution as:

	$X_i \in fg$	$X_i \in bg$
$X_i \in S$	α	$1-\alpha$
$X_i \notin S$	β	$1-\beta$

Table 1: **Binomial distribution for** $P(X_i|S)$

Where $\alpha \gg \beta$. Assuming the independence of pixels, the likelihood of observing the X as the foreground data given the codeword shape S is: $P(X|S) = \prod_{X_i \in image} P(X_i|S)$. The likelihood $P(X'|S')$ is estimated in the same way.

A codeword $\mathbf{C}$ includes both the human body shape S and its deformed shape S' in the next frame. The likelihood of observing the configuration $\mathbf{X}$ given a codeword $\mathbf{C} = (S, S')$ can then be estimated as:

$$P(\mathbf{X}|\mathbf{C}) = P(X, X'|S, S') = P(X|S)P(X'|S') \quad (1)$$

4 HMM for Action Recognition

Assume the observation contains the sequence of the foreground configurations as $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T)$, where $\mathbf{X}_t, t = 1 \dots T$ are defined as same as $\mathbf{X}$ in Section 3. To recognize the observed sequence as one of the learned actions, the slightly modified Hidden Markov Model (HMM) is applied. We define the states of HMM as the trained codewords. The probability of an observation generated from a state is obtained in the previous section as $P(\mathbf{X}|\mathbf{C})$. This probability is not obtained through the usual HMM learning scheme because here we know exactly by construction what the states are.

Assume there are K human actions we want to recognize: $\mathbf{H}_k, k = 1 \dots K$. In the training, each movelet has been assigned to one codeword. Therefore, for each training action k, we obtain a codeword sequence to represent its movelet sequence. Its HMM state (codeword) transition matrix A_k can be easily trained from this codeword sequence.

To recognize which action the observed sequence $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T)$ represents and what is its best representative codeword sequence, the following HMM solutions are applied. First, for each action hypothesis $k \in (1 \dots K)$, we apply the viberbi algorithm to search for the single best state (codeword) sequence which maximizes the following probability:

$$\begin{aligned} & q^*_{k1}, q^*_{k2}, \dots, q^*_{kT} \\ = & \max_{q_{k1}, q_{k2}, \dots, q_{kT}} P(q_{k1}, q_{k2}, \dots, q_{kT} | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T, A_k) \end{aligned}$$

Where $q_{kt} \in (\mathbf{C}_1 \ldots \mathbf{C}_N)$. The human action $\mathbf{H}_{k^*}$ that best represents the observed sequence is determined by:

$$k^* = \arg\max_k P(q^*_{k1}, q^*_{k2}, \ldots, q^*_{kT}, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T | A_k)$$

Correspondingly the sequence $q^*_{k^*1}, q^*_{k^*2}, \ldots, q^*_{k^*T}$ is the recognized most likely codeword sequence to represent the observed sequence $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T)$.

5 Experiments

5.1 Experimental Design for Codewords Training

The objective of training is to obtain the dataset of codewords from the movelets of training actions. The human body movelet is represented by the union of body parts and their motions. Therefore, we need to identify each body part in the image frames of training actions. For this purpose, we made a set of special clothes which has distinguishable color for each body part. Also our training subject wears black mask on the head. We make sure that any two neighboring parts have different colors and the two parts having same color are located far away from each other. Therefore each body part can be segmented by its color and location difference from the others and represented by a rectangle. A movelet is formed by stacking the rectangular shapes of all body parts visible on a pair of consecutive frames together. In a movelet, the shapes in both frames are relative to the same origin which is the center of the head position in the first frame. An example of a pair of training frames and the extracted movelets is shown in Fig.1. Given the movelets of all training actions, codewords are learned following the scheme proposed in Section 2.

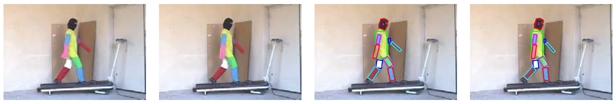

Figure 1: **Example of Training Images and Movelet**. Left two: two consecutive images in the training sequence. Right two: estimated rectangular shapes for both frames to form a movelet.

5.2 Matching of Codeword to Foreground Configuration

To estimate the likelihood $P(\mathbf{X}|\mathbf{C})$ that a foreground configuration is observed given a codeword , we need to efficiently search for the position in the image to locate the codeword so that it fits the foreground configuration best. Although the codewords are represented relative to the center of head, centroid is the most reliable point to detect a foreground object. Therefore, instead of searching for the head to locate the codeword, we match the codeword to the foreground configuration according to their centroids. For each codeword, we estimate the centroid of its shapes. Given a pair of test frames, the centroid of first frame's foreground data is computed. We then lock in the square of 49×49 pixels centered at this centroid as the searching region to place the codeword's centroid. Fig.2 shows an example of how log-likelihood varies with respect to the centroid placed in this region. For each codeword, we search in this region for the location where the observation log-likelihood of the first frame given the shape is maximized. The deformed shape of the codeword is correspondingly placed on the second frame. After locating the codeword, the log-likelihood of $P(\mathbf{X}|\mathbf{C})$ can be determined by Eq.1. We arbitrarily choose $\alpha = 0.9$ and $\beta = 0.1$.

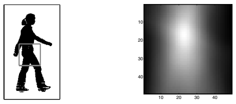

Figure 2: **Log-likelihood Varies With Centroid Position**. Left: the searching region to place the centroid of a codeword is the 49×49 square. Right: Log-likelihood $\log P(\mathbf{X}|\mathbf{C})$ varies with respect to the assumed centroid position in the searching region. It is gray-scaled. Brighter indicates higher log-likelihood.

In our experiment we actually search for the centroid location of a codeword on every third pixel in the searching region, which ends up with 16×16 possible searching locations. The computational time of such searching for one codeword is 0.012 seconds on Pentium 750MHz machine.

5.3 Periodic Action Recognition

In this experiment section we illustrate the recognition of a set of 9 view-based periodic actions: 3 different gaits (stepping at same place, walking and running) captured at 3 different view angles ($45°, 90°, 135°$ between the camera and subject direction). To make the action and view angle controllable, a treadmill was used on which all these periodic actions were performed. The image sequences were captured at 30 frames/sec.

Training the Actions

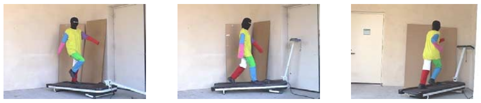

Figure 3: **Key Frames of the Training Periodic Actions**. 45° stepping,90° walking,135° running.

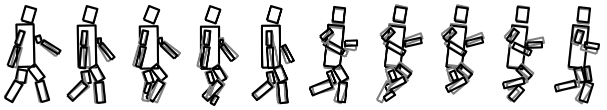

Figure 4: **Samples of Trained Codewords** (deformed shape S' is in gray, superimposed by the shape S in black). It shows samples of codewords that actions 90° walking (first 5 plots),90° running (last 5 plots) can pass, obtained from their HMM transition matrix.

The training actions were performed by 1 subject wearing the special colored clothes. For each training action, 1800 frames were collected and input to the codewords training algorithm. Some key frames are shown in Fig.3. Movelets are obtained from each pair of frames as described in section 5.1.

From this set of movelets, we first train $N = 270$ codewords. The performance of recognition with training of different number N of codewords is reported at the end of this section. As described in Section 2, the movelets are grouped according to their occlusion patterns. There are 15 occlusion patterns in this training set. The proportional number of codewords are trained from the group of movelets associated with each occlusion pattern.

Given the trained codewords and their associated movelets, we learn the transition matrix of HMM for each action. The transition matrix represents the possible temporal paths of codewords that an action can pass through. In Fig.4, we show a few samples of codewords that 90° walking and running actions pass.

Recognition of the Actions

6 subjects participated in the test experiments. A sequence of 1050 frames (35 seconds video) was captured for each of 9 actions performed by each subject. Foreground data were obtained by background subtraction. We answer three questions in this experiment: 1. how well can the actions be recognized? 2. how many frames are needed to recognize the action with certain accuracy? 3. how does the number of trained codewords influence the recognition performance?

Human vision doesn't need to observe the whole sequence (1050 frames) to recognize the action. To test how many frames are necessary for a good recognition, each sequence is split into segments of T frames long. The tail frames are ignored. We classify each segment into a learned action. Table 2 shows the confusion matrix at $T = 20$.

Act	45° S	90° S	135° S	45° W	90° W	135° W	45° R	90° R	135° R
45° S	66%	1%	42%						
90° S		99%							
135° S	34%		58%						
45° W				67%	3%	31%			
90° W					84%				
135° W				33%	13%	69%			
45° R							83%		57%
90° R								84%	
135° R							17%	16%	43%

Table 2: **Confusion Matrix for Periodic Action Recognition ($T = 20$)**

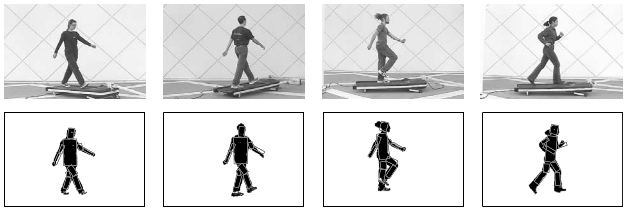

Figure 5: Top row: sample images of subject's 45° walking,135° walking,90° stepping,90° running. Bottom row: fit codewords.

We make the following two observations from table 2. First, the direction of the action is more difficult to recognize than the gait. In other words, we have no problem to tell if the action is stepping, walking or running, but have some difficulty in telling its direction. Secondly, we claim that 45° viewangle is basically not separable from 135° viewangle using only shape and motion cues. Fig.5 shows the examples of original images and best fit codeword of a few test actions. In this figure, although the first two actions, 45° and 135° walking, were performed by two different subjects, their individual foreground data are so similar to each other that they are represented by same codeword. Therefore 45° and 135° walking can be regarded as one action, so is 45° and 135° stepping, as well as 45° and 135° running.

The left two plots in Fig.6 demonstrate how the correct recognition rate varies as the segment length T varies from 1 to 40 for six actions (here we regard 45° and 135° actions as one already). From the plots, we observe that better performance is gained with longer segment length T. The action can be reliably recognized within about 20 frames, which is 0.66 seconds of action.

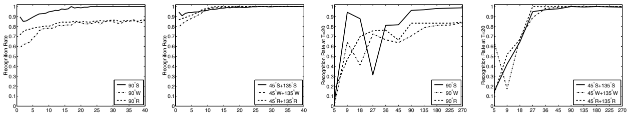

Figure 6: **Performance for Periodic Actions.** Left two: recognition rate vs. segment length T. Right two: recognition rate vs. number of trained codewords.

What is the appropriate number of codewords we should train to represent the space of movelets? This question can be answered by experiments. We train the different number $N = (5, 9, 18, 27, 36, 45, 90, 135, 180, 225, 270)$ of codewords from the training set of movelets, and repeat the recognition experiments with each group of trained codewords. The right two plots in Fig.6

Reach	$A000°$	$A045°$	$A090°$	$A135°$	$A180°$	$A225°$	$A270°$	$A315°$
$A000°$	90%							
$A045°$		100%						
$A090°$			100%					
$A135°$				100%				
$A180°$					100%			
$A225°$						100%		
$A270°$							100%	
$A315°$	10%							100%

Table 3: **Confusion Matrix for Nonperiodic Action Recognition**

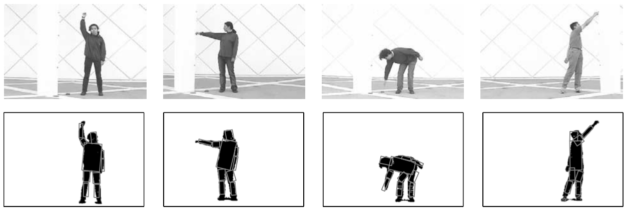

Figure 7: Top row: sample images of reaching actions for testing. Bottom row: fit codewords.

show how the recognition rate at segment length $T = 20$ varies with respect to different number of trained codewords. For all the actions, above 80% recognition rates are achieved with $N \geq 135$.

5.4 Nonperiodic Action Recognition

In this section we describe the experiments in modeling and recognizing nonperiodic reaching actions. The actions are reaching 8 different directions using right arm and captured with front view. From reaching the top, we name the actions counter clockwise as $A000°, A045°, A090°, A135°, A180°, A225°, A270°$ and $A315°$.

Training the Actions

The training actions were performed by 1 subject with the specially designed color clothes. The training subject repeated each reaching action 20 times which last 40 seconds. We learned 240 codewords first for this set of actions.

Recognition of the Actions

A set of 400 sequences of reaching different directions done by 5 subjects were captured for testing. Each sequence is about 60 frames (2 seconds video) long. For the nonperiodic action, the recognition is done by using all of the frames the sequence contains. The confusion matrix is shown in Table 3. We also train the different number $N = (4, 8, 16, 24, 32, 40, 80, 120, 160, 200, 240)$ of codewords for this set of actions, and repeat the recognition experiments with each group of trained codewords. The results are shown in Fig.8. For all the actions, above 80% recognition rates are achieved with $N \geq 120$ and 100% recognition rate are achieved for 6 out of 8 actions.

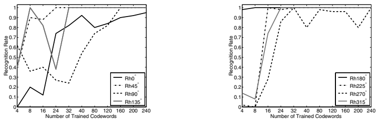

Figure 8: **Performance for Nonperiodic Actions.** Recognition Rate vs. Number of Trained Codewords.

6 Conclusion and Discussion

In this paper we proposed and tested an algorithm for the recognition of both human poses and action simultaneously in the image sequence. The experiments show above 80% recognition rate is achieved for all the test actions while over half of the actions can be recognized with the accuracy rate higher than 98%. Our fundamental idea can be extended to other types of action recognition, as long as the experimental setup for the training is properly designed. A possible future work that we may pursue is to recognize the actions from sequence without background subtraction.

References

[1] A. Bissacco, F. Nori, and S. Soatto, "Recognition of human gaits", In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 2001.

[2] C. Bregler, "Learning and recognizing human dynamics in video sequences", In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 568–574, 1997.

[3] J. Davis and A. Bobick, "The representation and recognition of action using temporal templates", In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 928–934, 1997.

[4] P. Felzenszwalb and D. Huttenlocher, "Efficient matching of pictorial structures", In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages II:66–73, 2000.

[5] I. Haritaoglu, D. Harwood, and L. Davis, "Ghost: a human body part labeling system using silhouettes", In *Proc. Int. Conf. Pattern Recognition*, pages 77–82, 1998.

[6] S. Ioffe and D. Forsyth, "Human tracking with mixtures of trees", *Proc. 8^{th} Int. Conf. Computer Vision*, pages I:690–695, 2001.

[7] R. Polana and R. Nelson, "Detecting activities", *Journal of Visual Communication and Image Representation*, 5(2):172–180, 1994.

[8] R. Rosales and S. Sclaroff, "Inferring body pose without tracking body parts", In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages II:721–727, 2000.

[9] Y. Song, X. Feng, and P. Perona, "Towards the detection of human motion", In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 810–817, 2000.

[10] Y. Yacoob and M. Black, "Parameterized modeling and recognition of activities", *Computer Vision and Image Understanding*, 73(2):232–247, 1999.

[11] J. Yamoto, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model", In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 379–385, 1992.