Human Action Recognition Using HDP by Integrating Motion and Location Information

Yasuo Ariki[1], Takuya Tonaru[2], and Tetsuya Takiguchi[1]

[1] Organization of Advanced Science and Technology, Kobe University,
1–1, Rokkodai, Nada, Kobe, Hyogo, Japan
{ariki,takigu}@kobe-u.ac.jp
[2] Graduate School of Engineering, Kobe University,
1–1, Rokkodai, Nada, Kobe, Hyogo, Japan
http://www.me.cs.scitec.kobe-u.ac.jp/index.html

Abstract. The method based on local features has an advantage that the important local motion feature is represented as bag-of-features, but lacks the location information. Additionally, in order to employ an approach based on bag-of-features, language models represented by pLSA and LDA (Latent Dirichlet Allocation) have to be applied to. These are unsupervised learning, but they require the number of latent topics to be set manually. In this study, in order to perform the LDA without specifying the number of the latent topics, and also to deal with multiple words concurrently, we propose unsupervised Multiple Instances Hierarchical Dirichlet Process MI-HDP-LDA by employing the local information concurrently. The proposed method, unsupervised MI-HDP-LDA, was evaluated for Weizmann dataset. The average recognition rate by LDA as conventional method was 61.8% and by the proposed method it was 73.7%, resulting in 11.9 points improvement.

Key words: Motion, location, action recognition, LDA, HDP, HDP-LDA

1 Introduction

Human action recognition is a challenging problem in computer vision. It can be applied to many applications such as surveillance, scene understanding, care monitoring, sport analysis, etc. In fact, for computers to support human works, they need to understand the human activities, so that human actions, the primitive units of human activities, become important information.

A lot of work has been done in recognizing human actions. Bobick[1] used motion energy images (MEI) and motion history images (MHI). Those shape descriptors represented information of human motion –"where" and "how". Grundmann[2] used 3D shape context extended into a temporal dimension. That method represented a human action as a histogram of 3D points by sampling shape of silhouette. Additionally, it increased the sampling density in the domain of fast moving body parts. Efros[3] used optical flow field for human figures at

each frame. Their methods presented the human body as a whole to understand human actions as concatenation of motion and pose.

In contrast to the above methods, local motion approaches represent human action as a set of distinguished local motion features. Laptev[4] proposed space-time interest points using Harris operator extended into temporal and adapted scale. Dollár[5] proposed cuboid with local motion descriptor at interest point detected using separable linear filters. Scovanner[6] used 3D SIFT descriptor extended into a temporal dimension for these interest point. These approaches are effective to characterize distinguished local motion included in the action. Moreover, since these features can be represented as a histogram, language models with unsupervised learning can be applied to the features. These approaches are called bag-of-words and the features obtained as a result are called word. Niebles[7] classified actions by applying pLSA that is one of the language models, and Wang[8] used Semi-LDA.

However, local motion approaches do not take the location information into consideration. Moreover, bag-of-words approach using language model requires the number of latent topics, corresponding to action classes, to be set manually. In this study, in order to perform the LDA (Latent Dirichlet Allocation) without specifying the number, and also to deal with multiple words concurrently in an unsupervised manner, we propose unsupervised Multiple Instances Hierarchical Dirichlet Process MI-HDP-LDA. MI-HDP-LDA is the model capable of generating words from the latent topics. Hence it can provide co-occurrence of words occurring simultaneously. Moreover, it can estimate the number of latent topics automatically by using Hierarchical Dirichlet Processes(HDP).

The rest of this paper is organized as follows. In section 2, our basic idea is described and in section 3, motion feature and location information are described. In section 4, Hierarchical Dirichlet Processes - Latent Dirichlet Allocation (HDP-LDA) is briefly described. In section 5, MI-HDP-LDA is proposed to deal with features occurring concurrently. In section 6, the experimental result is described for Weizmann dataset introduced in [9] to evaluate our algorithm. Section 7 is for conclusion of this paper.

2 Basic Idea

Our study is motivated by the conventional approaches which extract local features by detecting interest points. Our basic idea is to extract various types of information at interest points such as motion feature, location information and limbs parts, etc for understanding human actions.

This paper regards the motion in terms of information “where” and “how”. For “where”, the relative position in human region is used as location word and for “how”, the motion feature is used as motion word at the interest point. Niebles[7] method is employed as interest point detection algorithm and motion descriptor.

3 Features

In this section, we describe briefly the motion feature proposed by Dollár[5] and location information representing "where".

3.1 Motion Feature

Assuming a stationary camera or a process that can account for camera motion, separable linear filters are applied to the video to obtain the response function as follows,

$$\begin{aligned} R(x,y) =& \Big(I(x,y) * g(x,y;\sigma) * h_{ev}(t;\tau,\omega) \Big)^2 \\ &+ \Big(I(x,y) * g(x,y;\sigma) * h_{od}(t;\tau,\omega) \Big)^2, \end{aligned} \tag{1}$$

where $g(x,y;\sigma)$ is a 2D Gaussian smoothing kernel, applied only along the spatial dimensions, and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as follows,

$$\begin{aligned} h_{ev}(t;\tau,\omega) &= -\cos(2\pi t\omega) e^{-\frac{t^2}{\tau^2}}, \\ h_{od}(t;\tau,\omega) &= -\sin(2\pi t\omega) e^{-\frac{t^2}{\tau^2}}. \end{aligned} \tag{2}$$

The two parameters σ and τ correspond to the spatial and temporal scales of the filters respectively. To make the response function effective, $\omega = 4/\tau$ was employed.

This function detects any regions where complex motion is caused spatially. In fact, a region with complex motion can induce a strong response, but a region with simple translational motion will not induce a strong response. The spatial-temporal interest points are extracted around the local maxima of the response function. At each interest point, a spatial-temporal cube is extracted that contains the output of the response function. Its size is approximately six times the spatial and temporal scales along each dimension. To obtain a motion descriptor, the brightness gradients are computed at all the pixels in the cube and are concatenated to form a vector. Then PCA is applied to reduce the dimensionality of the descriptors.

In order to obtain the cluster prototypes, a k-means algorithm is applied to the descriptors. Then each descriptor is assigned a descriptor type by mapping it to the prototype. Therefore a collection of descriptors included in a video is represented as a histogram of the descriptor types. The descriptor types are called motion words.

3.2 Location Information

As shown in Fig.1, the human rectangle region is divided into $N \times M$ blocks. Each block indicates a relative position within a human rectangle region. The extremities of the motion such as arm and foot movement are extracted as the

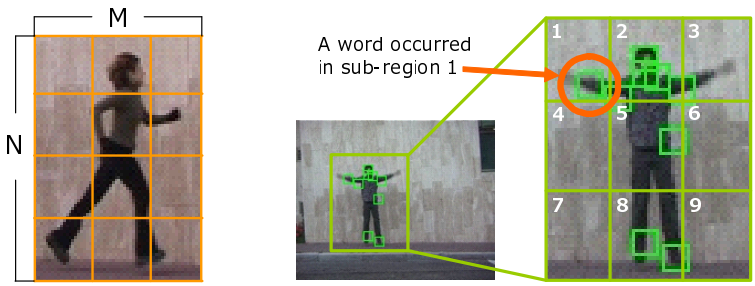


Fig. 1. (a)Human rectangle region is split into $N \times M$ blocks. (b)A motion word enclosed in orange circle also has a location word occurred in sub-region 1. This indicates that these words occurred at the interest point have motion feature and location information concurrently.

interest points and therefore they have two kinds of information, namely, motion feature and location information. The person detection is done manually in the experiment to exclude the detection errors, but it will be automatically performed by frame subtraction.

4 Hierarchical Dirichlet Processes - Latent Dirichlet Allocation

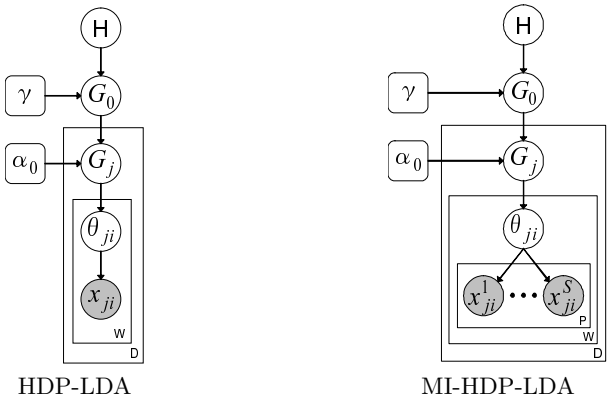


Fig. 2. Graphical representation of HDP-LDA model and MI-HDP-LDA model.

Our model is based on Hierarchical Dirichlet Processes - Latent Dirichlet Allocation (HDP-LDA) [10]. HDP-LDA is extended from LDA[11] by using multiple DPs. In contrast to LDA with a finite mixture model, HDP-LDA is an

infinite mixture model sharing topics across multiple DPs given an underlying base measure H. The graphical model of HDP-LDA is depicted in Fig.2 left.

Suppose we are given a collection D of video clips $\{1, \cdots, j, \cdots, J\}$. Video clip j has a collection of words $\{x_{j1}, \cdots, x_{ji}, \cdots, x_{jI}\}$ as described in the previous section, where x_{ji} is the i-th word in video clip j.

The global measure G_0 has a probability distribution decided by Dirichlet process(DP)[12] with concentration parameter γ and base probability measure H as follows,

$$G_0 \mid \gamma, H \sim \mathrm{DP}(\gamma, H). \tag{3}$$

DP is a process that a random probability measure is distributed as a Dirichlet distribution with concentration parameter and base probability measure. The random probability measure G_j for designated video clip j has a distribution decided by a Dirichlet process with concentration parameter α_0 and base probability measure G_0 under conditional independence given G_0,

$$G_j \mid \alpha_0, G_0 \sim \mathrm{DP}(\alpha_0, G_0). \tag{4}$$

Such hierarchical process of distributing a probability measure is Hierarchical Dirichlet Processes. For each video clip j, let $\theta_{j1}, \theta_{j2}, \cdots$ be independent and identically distributed random variables sampled from G_j. Each θ_{ji} is a topic corresponding to a single word x_{ji}. The likelihood is given by:

$$\theta_{ji} \mid G_j \sim G_j \tag{5}$$

$$x_{ji} \mid \theta_{ji} \sim F(\theta_{ji}), \tag{6}$$

where $F(\theta_{ji})$ denotes the probability distribution of the observation x_{ji} given θ_{ji}. Words are generated independently and distributed identically from the selected topic.

5 Multiple Instances HDP-LDA

HDP-LDA model generates a single word x_{ji} from the corresponding topic θ_{ji}, but not the multiple instances of the word concurrently such as motion word and location word. To solve this problem, we propose Multiple Instances HDP-LDA(MI-HDP-LDA) that allows multiple concurrent instances of the word.

MI-HDP-LDA can generate multiple instances $\boldsymbol{x}_{ji} = \{x_{ji}^1, \cdots, x_{ji}^S\}$ of the word from latent topic θ_{ji}. Each instances x_{ji}^s is generated as follows,

$$x_{ji}^s \mid \theta_{ji} \sim F_s(\theta_{ji}), \tag{7}$$

where $F_s(\cdot)$ is the distribution of x_{ji}^s given the latent topic θ_{ji}. Here S indicates the number of instances of the word i such as motion word and location word in the video clip j.

Next, we describe the Gibbs sampling scheme for MI-HDP-LDA in CRF (Chinese Restaurant Franchise) representation[10]. Basic scheme is exactly similar to HDP-LDA except it obtains the likelihood of generating $\boldsymbol{x}_{ji}$.

The variable $\boldsymbol{x}_{ji}$ is multiple instances of word i observed concurrently, so that $\boldsymbol{x}_{ji}$ is a vector with the size of S. Each $\boldsymbol{x}_{ji}$ is assumed to be generated based on a distribution $F(\theta_{ji})$. Let the factor θ_{ji} be associated with the table t_{ji} in CRF, i.e., let $\theta_{ji} = \psi_{jt_{ji}}$. The random variable ψ_{jt} is a topic k_{jt}; i.e., $\psi_{jt} = \phi_{k_{jt}}$. The prior over the parameters ϕ_k is H. Let $z_{ji} = k_{jt_{ji}}$ denote the topic associated with the observation $\boldsymbol{x}_{ji}$. We use the notation n_{jtk} to denote the number of customers in restaurant j at table t eating dish k, while m_{jk} denotes the number of tables in restaurant j serving dish k. Marginal counts are represented with dots.

Let $\boldsymbol{x} = \{\boldsymbol{x}_{ji} : \text{all } j, i\}$, $\boldsymbol{x}_{jt} = \{\boldsymbol{x}_{ji} : \text{all } i \text{ with } t_{ji} = t\}$, $\boldsymbol{t} = \{t_{ji} : \text{all } j, i\}$, $\boldsymbol{k} = \{k_{jt} : \text{all } j, t\}$, $\boldsymbol{z} = \{z_{ji} : \text{all } j, i\}$, $\boldsymbol{m} = \{m_{jk} : \text{all } j, k\}$, $\boldsymbol{\phi} = \{\phi_1, \cdots, \phi_K\}$. When a superscript is attached to a set of variables or a count, e.g., $\boldsymbol{x}^{-ji}$,$\boldsymbol{k}^{-jt}$ or $n_{jt\cdot}^{-ji}$, this means that the variable corresponding to the superscripted index is removed from the set or from the calculation of the count.

Sampling $\boldsymbol{t}$. The probability that t_{ji} takes on a previously used value t or new value t^{new} is given as follows;

$$p(\boldsymbol{x}_{ji}|\boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}) = \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} f_k^{-\boldsymbol{x}_{ji}}(\boldsymbol{x}_{ji}) + \frac{\gamma}{m_{\cdot\cdot} + \gamma} f_{k^{\text{new}}}^{-\boldsymbol{x}_{ji}}(\boldsymbol{x}_{ji}) \quad (8)$$

$$f_k^{-\boldsymbol{x}_{ji}}(\boldsymbol{x}_{ji}) = \int \prod_{s=1}^{S} F_s(x_{ji}^s|\phi_k) h(\phi_k) d\phi_k \quad (9)$$

$$f_{k^{\text{new}}}^{-\boldsymbol{x}_{ji}}(\boldsymbol{x}_{ji}) = \int \prod_{s=1}^{S} F_s(x_{ji}^s|\phi_{k^{\text{new}}}) h(\phi_{k^{\text{new}}}) d\phi_{k^{\text{new}}} \quad (10)$$

where $h(\phi_{k^{new}})$ is probability density function of the base probability measure H. The conditional distribution of t_{ji} is then obtained as follows;

$$p(t_{ji} = t|\boldsymbol{t}^{-ji}, \boldsymbol{k}) \propto \begin{cases} n_{jt\cdot}^{-ji} f_{k_{jt}}^{-\boldsymbol{x}_{ji}}(\boldsymbol{x}_{ji}) & \text{if } t \text{ is previously used,} \\ \alpha_0 f_{k^{\text{new}}}^{-\boldsymbol{x}_{ji}}(\boldsymbol{x}_{ji}) & \text{if } t = t^{\text{new}}. \end{cases} \quad (11)$$

If the sampled value of t_{ji} is t^{new}, we obtain a sample of $k_{jt^{\text{new}}}$ according to the following probability:

$$p(k_{jt^{\text{new}}} = k|\boldsymbol{t}, \boldsymbol{k}^{-jt^{\text{new}}}) \propto \begin{cases} m_{\cdot k} f_k^{-\boldsymbol{x}_{ji}}(\boldsymbol{x}_{ji}) & \text{if } k \text{ is previously used,} \\ \gamma f_{k^{\text{new}}}^{-\boldsymbol{x}_{ji}}(\boldsymbol{x}_{ji}) & \text{if } k = k^{\text{new}}. \end{cases} \quad (12)$$

Sampling $\boldsymbol{k}$. Since k_{jt} actually changes the component membership of all data items in table t, the likelihood obtained by setting $k_{jt} = k$ is given by

$f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt})$, so that the conditional probability of k_{jt} is obtained as follows;

$$p(k_{jt}=k|\boldsymbol{t},\boldsymbol{k}^{-jt}) \propto \begin{cases} m_{\cdot k}^{-jt} f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) & \text{if } k \text{ is previously used,} \\ \gamma f_{k^{\text{new}}}^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) & \text{if } k = k^{\text{new}}, \end{cases} \tag{13}$$

$$f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) = \prod_{i:t_{ji}=t} \int \prod_{s=1}^{S} F_s(x_{ji}^s|\phi_k) h(\phi_k) d\phi_k, \tag{14}$$

$$f_{k^{\text{new}}}^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) = \prod_{i:t_{ji}=t} \int \prod_{s=1}^{S} F_s(x_{ji}^s|\phi_{k^{\text{new}}}) h(\phi_{k^{\text{new}}}) d\phi_{k^{\text{new}}}. \tag{15}$$

In recognition, topics of each $\boldsymbol{x}_{ji}$ are calculated using Gibbs sampling with $F(\theta)$ obtained in learning. Given a test video j' and words $\boldsymbol{x}_{j'i}$, the corresponding topic $\theta_{j'i}$ is computed and the topic histogram of test video j' is obtained as follows,

$$\text{hist}(\theta) += \theta_{j'i}, \tag{16}$$

$$k = \max(\text{hist}(\theta)). \tag{17}$$

An action label recognized for test video j' is the maximization topic k of the histogram.

6 Experiments

The proposed method was evaluated for Weizmann dataset which includes 10 motion classes such as jump, run, ship and walk. The total number of movies included in the database was 92. We employed leave-one-out cross validation as evaluation method. The following four experiments were conducted for the evaluation.

- Exp. 1: motion + LDA
  - LDA was evaluated using only motion word as a baseline method.
- Exp. 2: motion + HDP-LDA
  - HDP-LDA was evaluated using only motion word for comparison with Exp.1.
- Exp. 3: motion + loation + HDP-LDA
  - HDP-LDA was evaluated using motion word and location word to compare it with MI-HDP-LDA.
- Exp. 4: motion + loation + MI-HDP-LDA
  - The proposed method was evaluated using motion word and location word.

At first, LDA was evaluated using motion word as a baseline. Though LDA generates the prior distribution by using the Dirichlet distribution as well as HDP-LDA, it can not estimate the number of latent topics automatically as HDP can do.

In the motion word parameters, cuboid size was $15 \times 15 \times 15$ and codebook size was 1000. In the location word, the number of blocks in human region size was 10×13. The response parameter τ was set to 5 and PCA reduced the dimension to 779. The number of the latent topics was set to 10 manually for LDA.

As a result of the experiment 1, the recognition rate was 61.8% and the confusion matrix is shown in Fig.4(a). It recognized the motions of bend, jack, pjump and wave1 excellently, but the motions of jump, side and skip were considerably confused. The reason of the confusion will be attributed to their similar movements of the body except for the legs. The motions of wave1 and wave2 had been learned as the same action class by unsupervised LDA without location information, because the subject waves right hand only in the motion wave1 and waves both hands in the motion wave2. Therefore it classified them into the same action class in the recognition.

As a result of the experiment 2, the recognition rate was 64.9% and the confusion matrix is shown in Fig.4(b). The average number of classes automatically estimated was 16.33. It recognized bend and jack, etc. excellently, but the motion of Jump and wave2 was confused as experiment 1.

As a result of the experiment 3, the recognition rate was 64.0% and the confusion matrix is shown in Fig.4(c). The average number of classes estimated was 14.33. This experiment was carried out using both the motion words and the location words for HDP-LDA. In this experiment, it was assumed that the both words were not concurrently occurred but were independently generated, and the word of the location information was simply added. This experiment was carried out to compare with MI-HDP-LDA in terms of information concurrency. The same likelihood $F_s(\theta)$ was used for the same word in HDP-LDA and MI-HDP-LDA.

Finally, experiment 4 was carried out for MI-HDP-LDA using both the motion words and the location words. The recognition rate was improved up to the highest score 73.7%. The average number of topics estimated was 15.78. The confusion matrix is shown in Fig.4(d). Especially, in unsupervised MI-HDP-LDA, wave1 and wave2 were automatically learned as different motion owing to location information of the motion, therefore they were classified into different motion classes in the recognition.

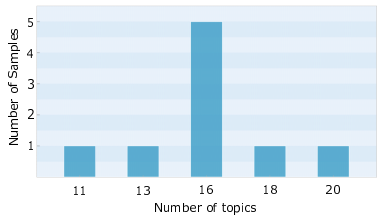
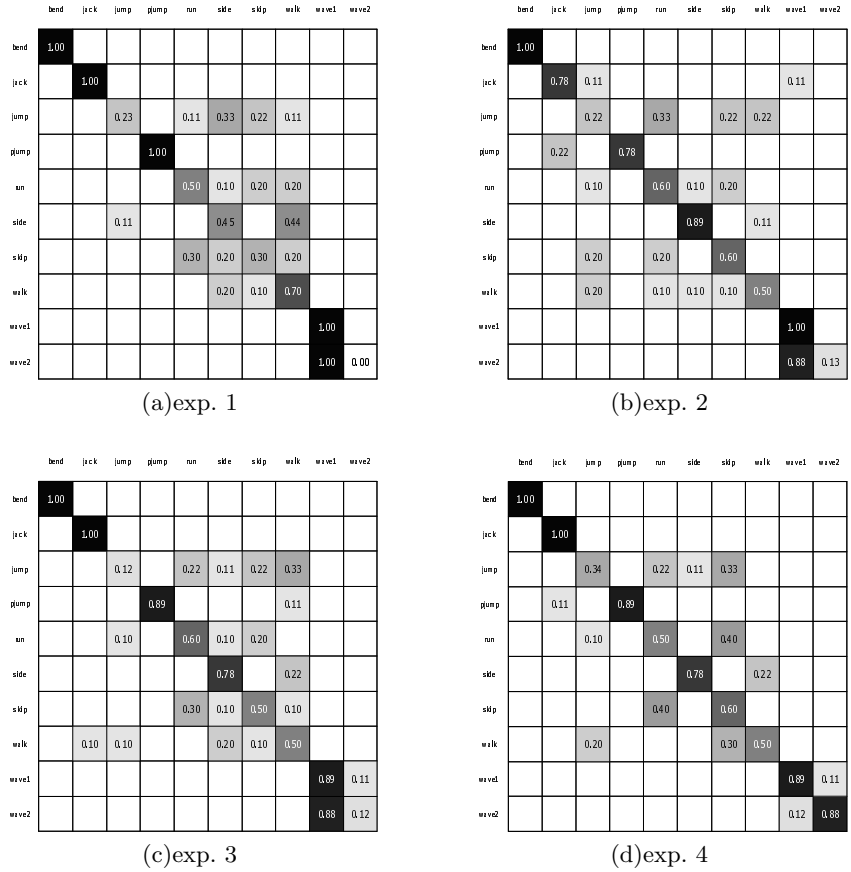


Fig. 3. Number of topics in experiment 4 (MI-HDP-LDA).



(a)exp. 1 (b)exp. 2

(c)exp. 3 (d)exp. 4

Fig. 4. Confusion matrices computed in respective experiment for Weizmann Dataset(%).

Next, the number of latent topics estimated by HDP is described. There were about 15 latent topics at average with some variation over the experiment 2, 3 and 4. Fig.3 shows the number of topics sampled in experiment 4.

Weizmann Dataset has two kinds of motions in jump, run, side, skip and walk: motions toward right or left directions. If these actions are separately counted, the number of actions included in Weizmann Dataset becomes 15.

It can be confirmed that the extended number of actions 15 almost coincides with the number of latent topics estimated in the experiment. The number of latent topics is the number of mixtures with the highest likelihood and estimated through the experiment, depending on the training data.

7 Conclusion

In this paper, a new unsupervised learning method MI-HDP-LDA has been proposed to deal with motion feature and location information concurrently in the motion recognition task. This method can also estimate the number of latent topics included in the training data automatically owing to the HDP (Hierarchical Dirichlet Processes).

In the experiments of motion learning and recognition for Weismann Dataset, LDA showed 61.8% recognition rate using only motion information. The proposed MI-HDP-LDA achieved 73.7% recognition rate, resulting in 11.9 points improvement.

Future work will be the incorporation of the various information such as "what" in addition to "where" and "how" which this paper pays attention to. Pose information of the limbs will be also important in learning and recognizing of the motion.

References

1. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(3) (2001) 257–267
2. Grundmann, M., Meier, F., Essa, I.: 3d shape context and distance transform for action recognition. In: International Conference on Pattern Recognition. (2008) 1–4
3. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision. (2003) 726–733
4. Laptev, I., Lindeberg, T.: Space-time interest points. In: IEEE International Conference on Computer Vision. (2003) 432–439
5. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: In VS-PETS. (2005) 65–72
6. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of international conference on Multimedia. (2007) 357–360
7. Niebles, J., Wang, H., Li.Fei-Fei: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79**(3) (2008) 299–318
8. Wang, Y., Sabzmeydani, P., Mori, G.: Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In: Workshop on Human Motion 2007. (2007) 240–254
9. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: IEEE International Conference on Computer Vision. (2005) 1395–1402
10. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the American Statistical Association **101**(476) (2006) 1566–1581
11. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
12. Ferguson, T.: A bayesian analysis of some nonparametric problems. Annals of Statistics **1**(2) (1973) 209–230