Learning Probabilistic Semantic Network of Object-Oriented Action and Activity

Masayasu Atsumi

Department of Information Systems Science, Faculty of Engineering,
Soka University, 1-236 Tangi, Hachioji, Tokyo 192-8577, Japan
masayasu.atsumi@gmail.com

Abstract. This paper proposes a method of learning probabilistic semantic networks which represent visual features and semantic features of object-oriented actions and their contextual activities. In this method, visual motion feature classes of actions and activities are learned by an unsupervised Incremental Probabilistic Latent Component Analysis (I-PLCA) and these classes and their semantic tags in the form of case triplets are integrated into probabilistic semantic networks to visually recognize and verbally infer actions in the context of activities. Through experiments using video clips captured with the Kinect sensor, it is shown that the method can learn, recognize and infer object-oriented actions in the context of activities.

Keywords: action recognition, activity recognition, probabilistic learning, semantic network, probabilistic inference.

1 Introduction

It is necessary for a human support robot to understand what a person is doing in everyday living environment. In human motion in everyday life, there is a lot of motion that interacts with objects, which is referred to as an "object-oriented motion" in this research. The meaning of an object-oriented motion is determined not only a motion itself but also an object with which the motion interacts and this view corresponds to an affordance in which motion is dependent on object perception. In addition, each motion is performed in a context which is defined by a sequence of motions and a certain motion frequently occurs in some context and rarely occurs in other contexts. For example, a motion using a fork is frequently observed in a context of eating meals. In this research, each object-oriented motion is referred to as an action and a sequence of actions is referred to as an activity and it is assumed that an activity gives a context of an action and promotes action recognition. This assumption is consistent with findings that context improves category recognition of ambiguous objects in a scene [1] and requires an extension to action recognition of several methods [2,3] which incorporate context into object categorization. Though object-oriented actions and activities can be clustered into motion classes according to visual motion features and their semantic features can be labeled by using motion

G. Agre et al. (Eds.): AIMSA 2014, LNAI 8722, pp. 1–12, 2014.

labels and their target labels, motion classes and their labels do not have one-to-one correspondence. Therefore, in this research, motion classes are labeled with target synsets and motion synsets of case triplets in a form of "⟨target synset, case, motion synset⟩" and motion classes and synsets are probabilistically linked in probabilistic semantic networks of actions and activities, where a synset is a synonymous set of the WordNet [4] which represents a primitive semantic feature. In addition, contextual relationship between actions and activities is acquired as co-occurrence between them.

This paper proposes a method of learning probabilistic semantic networks which represent visual features and semantic features of object-oriented actions and their contextual activities. It also provides a probabilistic recognition and inference method of actions and activities based on probabilistic semantic networks. The main characteristics of the proposed method are the following: (1) visual motion feature classes of actions and activities are learned by an unsupervised "Incremental Probabilistic Latent Component Analysis (I-PLCA)" [5], (2) visual feature classes of motion and synsets of case triplets are integrated into probabilistic semantic networks to visually recognize and verbally infer actions and activities, and (3) actions are inferred in the context of activities through acquired co-occurrence between them.

As for related work, Kitani et al. [6] proposed a categorization method of primitive actions in video by leveraging related objects and relevant background features as context. Yao et al. [7] proposed a mutual context model to jointly recognize objects and human poses in still images of human-object interaction. Also in [8], they proposed a model to identify different object functionalities by estimating human poses and detecting objects in still images of different types of human-object interaction. One difference between our method and these existing methods is that our method not only uses an action and its target object as mutual context but also uses activities as context of actions. Another difference is that our method probabilistically infers actions and activities by using different semantic features linked with different visual motion features in probabilistic semantic networks.

2 Proposed Method

2.1 Overview

Human motion is captured as a temporal sequence of three-dimensional joint coordinates of a human skeleton, which can be captured with the Microsoft Kinect sensor. Since this research focuses on object-oriented motions of hands, a temporal sequence of three-dimensional coordinates of both hand joints relative to a shoulder center joint are computed from human skeleton data.

A motion feature of both hands is computed from a temporal sequence of relative three-dimensional coordinates of both hand joints by the following procedure. First, relative three-dimensional coordinates of both hand joints are spatially-quantized at a certain interval and a temporal sequence of quantized coordinates and their displacement are obtained as a motion representation. Next,

actions are manually segmented from a temporal sequence of quantized coordinates and their displacement and they are semantically annotated with case triplets. An activity is also segmented as a sequence of actions and is semantically annotated with a case triplet. Then, for a temporal sequence of quantized coordinates and their displacement of an action, a motion feature of the action is computed as a motion histogram by firstly dividing the space around a shoulder center into modestly-sized regions, secondly calculating a displacement histogram of each region and lastly combining them into one histogram. A motion histogram of an activity is also calculated in the same way.

A probabilistic semantic network of actions is learned from a set of their motion histograms with case triplets. First of all, a set of action classes is generated from a set of motion histograms by the probabilistic clustering method I-PLCA. Here, an action class represents a motion feature of the action. Then, a probabilistic semantic network is generated as a network whose nodes consist of action classes and synsets of case triplets and are linked by joint probabilities of these classes and synsets. A probabilistic semantic network of activities is also learned from a set of their motion histograms with case triplets as a network whose nodes consist of activity classes and synsets of case triplets and are linked by joint probabilities of these classes and synsets. Here, an activity class is generated by the I-PLCA and represents a motion feature of the activity. For probabilistic semantic networks of actions and activities, co-occurrence between an action and an activity is obtained by calculating pointwise mutual information of their case triplets. A pair of probabilistic semantic networks of actions and activities with co-occurrence between them is referred to as the ACTNET in this paper. Fig. 1 shows an example of an ACTNET.

For a sequence of motion histograms of actions, actions and an activity are sequentially guessed by recognizing action and activity classes and inferring synsets of case triplets of them. First of all, an action class is recognized with the degree of confidence for a motion histogram of each action and at the same time an activity class is recognized with the degree of confidence for a sum of motion

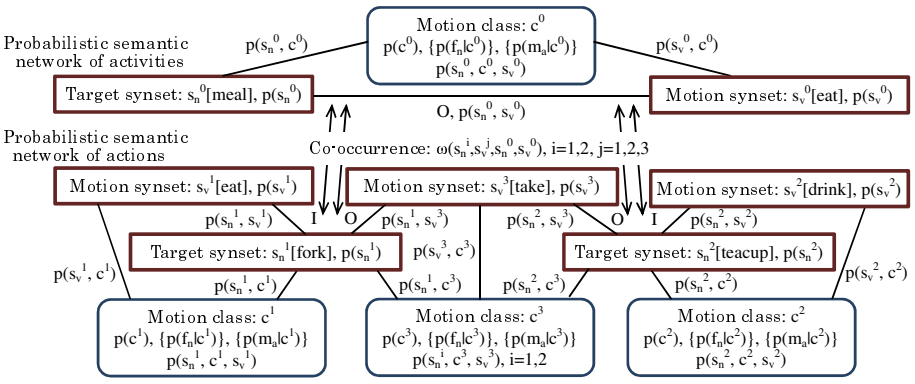

Fig. 1. An example of an ACTNET (Symbols in the figure are explained in the text.)

histograms of an action sequence. Then, synsets of case triplets of the actions and the activity are inferred from these classes and co-occurrence between them on probabilistic semantic networks.

2.2 Motion Feature of Action and Activity

Let $p^l = (p^l_x, p^l_y, p^l_z)$ and $p^r = (p^r_x, p^r_y, p^r_z)$ be relative quantized three-dimensional coordinates of left and right hands and let $d^l = (d^l_x, d^l_y, d^l_z)$ and $d^r = (d^r_x, d^r_y, d^r_z)$ be their displacement respectively. Here, l represents a left hand, r represents a right hand and the displacement is given by the difference of quantized coordinates between two successive frames. Let $\langle s_n[w_n], r, s_v[w_v]\rangle$ be a case triplet which is used to annotate a temporal sequence of quantized coordinates and their displacement of an action or an activity, where w_n is a noun which represents a target of motion and s_n is its synset, w_v is a verb which represents motion and s_v is its synset, and r is a case notation. Here, a synset is given by a synonymous set of the WordNet [4] and a case is currently one of the objective case (O), the instrumental case (I) and the locative case ($L[at|\ inside|\ around|\ above|\ below|\ beyond|\ from|\ to]$). For a temporal sequence of quantized coordinates and their displacement of an action, a case triplet of an activity which includes the action is also given in addition to a case triplet of the action. Then, for a motion $m = \{((p^l, d^l), (p^r, d^r))_t\}$, a motion histogram is constructed to represent a motion feature of both hands around a shoulder center as follows. Let B be a set of modestly-sized regions which is obtained by dividing the space around a shoulder center and let $|B|$ be the number of regions. For each region $b \in B$, a motion sub-histogram is computed for a set of coordinate and displacement data $\{(p^l, d^l)\}$ and $\{(p^r, d^r)\}$ whose coordinate p^l or p^r is located in the region b. A motion sub-histogram has 27 bins each bin of which corresponds to whether displacement is positive, zero or negative along x-, y-, and z-axes and is counted up according to values of displacement d^l and d^r. A motion histogram $h(m)$ is constructed as a $|B|$-plex histogram by combining these sub-histograms into one histogram so that the size of $h(m)$ is $27 \times |B|$.

2.3 Learning Probabilistic Semantic Network

An ACTNET is learned through generating probabilistic semantic networks of actions and activities followed by calculating co-occurrence between actions and activities. A probabilistic semantic network is generated from a set of motion histograms of actions or activities with their case triplets. Let $h(m_a) = [h_{m_a}(f_1), \ldots, h_{m_a}(f_{|F|})]$ be a motion histogram of a motion m with a case triplet a and let $H = \{h(m_a)\}$ be a set of motion histograms with case triplets. Here, $f_i \in F$ is a bin of a histogram and the size of a histogram is $|F| = 27 \times |B|$. A probabilistic semantic network is obtained through generation of motion classes by clustering a set of motion histograms H using I-PLCA [5] and derivation of a probabilistic network whose nodes consist of motion classes and synsets of case triplets and they are linked by joint probabilities of these classes and synsets.

The problem to be solved for generating motion classes is estimating probabilities $p(m_a, f_n) = \sum_c p(c)p(m_a|c)p(f_n|c)$, namely, a class probability distribution $\{p(c)|c \in C\}$, instance probability distributions $\{p(m_a|c)|m_a \in M \times A, c \in C\}$, probability distributions of class features $\{p(f_n|c)|f_n \in F, c \in C\}$, and the number of classes $|C|$ that maximize the following log-likelihood

$$L = \sum_{m_a} \sum_{f_n} h_{m_a}(f_n) \log p(m_a, f_n) \tag{1}$$

for a set of motion histogram $H = \{h(m_a)\}$, where C is a set of motion classes, M is a set of motions, A is a set of case triplets and m_a is a motion m with a case triplet a, that is, an instance of an action or an activity. These probability distributions and the number of classes are estimated by the tempered EM algorithm with subsequent class division. The process starts with one or a few classes, pauses at every certain number of EM iterations less than an upper limit and calculates the following dispersion index

$$\delta_c = \sum_{m_a} \left(\left(\sum_{f_n} \left| p(f_n|c) - \frac{h_{m_a}(f_n)}{\sum_{f'_n} h_{m_a}(f'_n)} \right| \right) \times p(m_a|c) \right) \tag{2}$$

for $\forall c \in C$. Then a class whose dispersion index takes the maximum value among all classes is divided into two classes. Let c^* be a source class to be divided and let c_1 and c_2 be target classes after division. Then, for a motion $m_a^* = \arg\max_{m_a}\{p(m_a|c^*)\}$ which has the maximum instance probability and its motion histogram $h(m_a^*) = [h_{m_a^*}(f_1), ..., h_{m_a^*}(f_{|F|})]$, one class c_1 is set by specifying a probability distribution of a class feature, an instance probability distribution and a class probability as

$$p(f_n|c_1) = \frac{h_{m_a^*}(f_n) + \kappa}{\sum_{f_{n'}} (h_{m_a^*}(f_{n'}) + \kappa)}, \quad \forall f_n \in F \tag{3}$$

$$p(m_a|c_1) = \begin{cases} p(m_a^*|c^*) & \cdots m_a = m_a^* \\ \frac{1-p(m_a^*|c^*)}{|M|-1} & \cdots \forall m_a (m_a \neq m_a^*) \end{cases} \tag{4}$$

$$p(c_1) = \frac{p(c^*)}{2} \tag{5}$$

respectively where κ is a positive correction coefficient. Another class c_2 is set by specifying a probability distribution of a class feature $\{p(f_n|c_2)|f_n \in F\}$ at random, an instance probability distribution $\{p(m_a|c_2)\}$ as 0 for m_a^* and an equal probability $\frac{1}{|M|-1}$ for $\forall m_a (m_a \neq m_a^*)$, and a class probability as $p(c_2) = \frac{p(c^*)}{2}$. This class division process is continued until dispersion indexes or class probabilities of all the classes become less than given thresholds. The temperature coefficient of the tempered EM is set to 1.0 until the number of classes is fixed and after that it is gradually decreased according to a given schedule until the EM algorithm converges and all the probability distributions are determined.

A probabilistic semantic network is derived from a class probability distribution $\{p(c)|c \in C\}$ and instance probability distributions $\{p(m_a|c)|m_a \in M \times A, c \in C\}$ associated with motion classes. The network nodes consist of motion class nodes and synset nodes. A synset node is derived from a target synset s_n or a motion synset s_v of a case triplet $\langle s_n[w_n], r, s_v[w_v]\rangle$ which is given by an instance m_a of an action or an activity where $a = \langle s_n[w_n], r, s_v[w_v]\rangle$. A motion class node is derived from a motion class $c \in C$ and has a class probability $p(c)$, a probability distribution of a class feature $\{p(f_n|c)|f_n \in F\}$, an instance probability distribution $\{p(m_a|c)|m_a \in M \times A\}$ and a set of joint probabilities each of which is a joint probability with a target synset s_n and a motion synset s_v of a case triplet that annotates the class c and is given by the expression (6)

$$p(s_n, c, s_v) = p(c) \quad \times \sum_{a=\langle s_n[*],*,s_v[*]\rangle} p(m_a|c) \tag{6}$$

where $*$ represents any word or any case. The network link between two nodes of a motion class c and a target synset s_n has a joint probability $p(s_n, c)$. The network link between two nodes of a motion class c and a motion synset s_v has a joint probability $p(s_v, c)$. The network link between two nodes of a target synset s_n and a motion synset s_v has a joint probability $p(s_n, s_v)$. The network nodes of a target synset s_n and a motion synset s_v have probabilities $p(s_n)$ and $p(s_v)$ respectively. These probabilities are computed by the expressions (7).

$$\begin{gathered} p(s_n, c) = \textstyle\sum_{s_v} p(s_n, c, s_v),\ p(s_v, c) = \sum_{s_n} p(s_n, c, s_v) \\ p(s_n, s_v) = \textstyle\sum_c p(s_n, c, s_v) \\ p(s_n) = \textstyle\sum_c p(s_n, c),\ p(s_v) = \sum_c p(s_v, c) \end{gathered} \tag{7}$$

In addition, a noun w_n of a target synset s_n and a verb w_v of a motion synset s_v are set to the target synset node and the motion synset node respectively.

Co-occurrence between actions and activities is computed between a pair of a target synset s_n and a motion synset s_v of an action and a pair of a target synset s_n^0 and a motion synset s_v^0 of an activity when the action has a case triplet $\langle s_n[w_n], r, s_v[w_n]\rangle$ and is included in the activity with a case triplet $\langle s_n^0[w_n^0], r^0, s_v^0[w_n^0]\rangle$. Let $p(s_n, s_v)$ be a joint probability of a target synset s_n and a motion synset s_v of an action and let $p(s_n^0, s_v^0)$ be a joint probability of a target synset s_n^0 and a motion synset s_v^0 of an activity. Then, co-occurrence between them is defined by the expression (8)

$$\omega(s_n, s_v, s_n^0, s_v^0) = \log \frac{p(s_n, s_v, s_n^0, s_v^0)}{p(s_n, s_v)p(s_n^0, s_v^0)} \tag{8}$$

where a joint probability $p(s_n, s_v, s_n^0, s_v^0)$ is calculated from action instances according to the expression (9)

$$p(s_n, s_v, s_n^0, s_v^0) = \sum_c \Big(p(c) \quad \times \sum_{a=\langle s_n[*],*,s_v[*]\rangle @ \langle s_n^0[*],*,s_v^0[*]\rangle} p(m_a|c)\Big) \tag{9}$$

where $a = \langle s_n[*], *, s_v[*]\rangle @ \langle s_n^0[*], *, s_v^0[*]\rangle$ means that an action m_a has a case triplet which matches a pattern $\langle s_n[*], *, s_v[*]\rangle$ and its contextual activity has a case triplet which matches a pattern $\langle s_n^0[*], *, s_v^0[*]\rangle$.

2.4 Recognition and Inference of Action and Activity

An ACTNET is used to recognize and infer a sequence of actions and an activity for a given sequence of motion histograms of actions. First of all, for each motion histogram, a motion class of an action is recognized with the degree of confidence and at the same time a motion class of an activity is recognized with the degree of confidence for a sequential sum of motion histograms of an action history. Then, synsets of case triplets of the actions and the contextual activity are inferred from these classes and co-occurrence between actions and activities.

Motion classes of an action and an activity are respectively recognized for a motion histogram and a sequential sum of motion histograms of an action history. Let $h(m) = [h_m(f_1), \ldots, h_m(f_{|F|})]$ be a motion histogram or a sum of motion histograms and let $\hat{h}(m) = [\hat{h}_m(f_1), \ldots, \hat{h}_m(f_{|F|})]$ be a distribution of it. Then, a motion class is obtained through calculating similarity between probability distributions of class features $\{p(f_n|c)|f_n \in F, c \in C\}$ and the distribution $\hat{h}(m)$ by the expression (10) and selecting the most similar class of all the classes C.

$$\beta(c, m) = 1 - \frac{\sum_{f_n} |p(f_n|c) - \hat{h}_m(f_n)|}{2} \tag{10}$$

This similarity provides the degree of confidence of a motion class.

For a selected class, a target synset and a motion synset are inferred with their degrees of confidence. Let c be a motion class of an action or an activity and let β be its degree of confidence. Then, a target synset s_n, a motion synset s_v and a pair of them are respectively inferred with their degrees of confidence $p(s_n|c) \times \beta$, $p(s_v|c) \times \beta$ and $p(s_n, s_v|c) \times \beta$ by following links from a node of the motion class c to adjacent synset nodes of s_n and s_v and retrieving probabilities maintained in those nodes and links of an ACTNET. When additional information about either a target synset or a motion synset is given, a motion synset or a target synset is respectively inferred with the degree of confidence $p(s_v|c, s_n) \times \beta$ or $p(s_n|c, s_v) \times \beta$ on an ACTNET. By introducing co-occurrence in the above inference, an action and an activity are interdependently inferred with the degree of confidence. Let c and β be a motion class of an action and its degree of confidence and let c^0 and β^0 be a motion class of an activity and its degree of confidence. Then, a pair of a target synset s_n and a motion synset s_v of the action and a pair of a target synset s_n^0 and a motion synset s_v^0 of the activity are inferred with the degree of confidence $\beta(s_n, s_v, s_n^0, s_v^0|c, c^0)$ that is calculated by the expression (11) which incorporates co-occurrence between them

$$\beta(s_n, s_v, s_n^0, s_v^0|c, c^0) = p(s_n, s_v|c) \times p(s_n^0, s_v^0|c^0) \times \frac{(\beta + \beta^0)}{2} + \lambda \times \omega(s_n, s_v, s_n^0, s_v^0) \tag{11}$$

where λ is a co-occurrence coefficient. When a pair of synsets (s_n^0, s_v^0) of the activity is fixed to (s_n^*, s_v^*) by additional information, a pair of synsets (s_n, s_v) of the action is inferred with the degree of confidence $\beta(s_n, s_v, s_n^*, s_v^*|c, c^0)$.

Table 1. Case triplets for activities and actions in a small data set

Activity	Action
<07573696-n[meal], O, 01166351-v[eat]>	<03383948-n[fork],O,01216670-v[take]> <03383948-n[fork],I,01166351-v[eat]> <03383948-n[fork],O,01494310-v[put]> <04398044-n[teapot],O,01216670-v[take]> <04398044-n[teapot],I,02070296-v[pour]> <04398044-n[teapot],O,01494310-v[put]> <04397452-n[teacup],O,01216670-v[take]> <04397452-n[teacup],I,01170052-v[drink]> <04397452-n[teacup],O,01494310-v[put]>
<03561345-n[illustration], O, 01684663-v[paint]>	<06415419-n[notebook],O,02311387-v[take-out]> <06415419-n[notebook],O,01346003-v[open]> <06415419-n[notebook],O,01291941-v[close]> <06415419-n[notebook],O,01308381-v[put-back]> <03908204-n[pencil],O,01216670-v[take]> <03908204-n[pencil],I,01684663-v[paint]> <03908204-n[pencil],O,01494310-v[put]>

3 Experiments

3.1 Experimental Framework

Experiments were conducted to evaluate learning, recognition and inference of object-oriented actions and activities on ACTNETs by using video clips captured with the Kinect Sensor. Relative three-dimensional coordinates of both hands are interpolated to about $30fps$ and are quantized at an interval of $1cm$. The space around a shoulder center is divided as follows for constructing a motion histogram. The front and the side in the vicinity of the body are both divided into 9 regions the size of which is $30cm$ on a side. The front and the side outside of these regions are respectively divided into 9 and 8 major regions and the back of the body is covered by just one region. As a result, the number of regions are 36 and the size of a motion histogram is $972(= 27 \times 36)$.

Two data sets of video clips were prepared for experiments. One is a small data set which is used to illustrate the basic capability of the proposed method in the experiment 1 and the other is a larger data set which is used to evaluate the performance of the proposed method in the experiment 2. A small data set contains 2 activities which are annotated by case triplets <07573696-n[meal],O,01166351-v[eat]> and <03561345-n[illustration],O,01684663-v[paint]>. The former activity "eating a meal" contains 3 objects and 9 object-oriented actions and the latter activity "painting an illustration" contains 2 objects and 7 object-oriented actions. The total number of actions is 16. Table 1 shows case triplets for activities and actions in this data set. Fig. 2 shows examples of motion quantization, that is, quantized coordinates and their displacement of both hands of some actions in this data set. A large data set contains 4 activities, 10 objects and

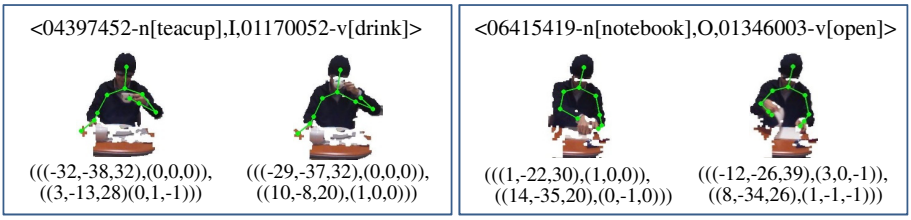

Fig. 2. Examples of quantized motion of both hands in actions

Table 2. Activities and action sequences in a large data set

Activity	Action sequence
<clothes,O,wear>	<necktie,O,take>,<necktie,O,tie>,<jacket,O,take> <jacket,O,wear>,<scarf,O,take>,<scarf,O,wind>
<meal,O,eat>	<fork,O,take>,<fork,I,eat>,<fork,O,put> <teapot,O,pick-up>,<teapot,I,pour>,<teapot,O,put> <teacup,O,pick-up>,<teacup,I,drink>,<teacup,O,put>
<desk,O,clean-up>	<notebook-computer,O,close>,<mouse,O,put-back> <book,O,close>,<book,O,put-back> <mop,O,take>,<mop,I,wipe-up>
<report,O,write>	<notebook-computer,O,open>,<book,O,take> <book,O,open>,<book,O,turn>,<teacup,O,pick-up> <teacup,I,drink>,<teacup,O,put> <mouse,O,operate>,<notebook-computer,I,input>

27 object-oriented actions in total. Table 2 shows these activities and object-oriented action sequences in the abbreviated form without synsets. Fig. 3 shows snapshots of object-oriented action sequences in two activities in this data set.

The parameters were set as follows. In the I-PLCA, a threshold of the dispersion index for class division was 0.1, a threshold of the class probability for class division was 0.1 for a small data set and 0.05 for a large data set respectively, and a correction coefficient κ in the expression (3) was 1.0. In the tempered EM, a temperature coefficient was decreased by multiplying it by 0.95 at every 20 iterations until it became 0.8. A co-occurrence coefficient λ between an action and an activity in the expression (11) was set to 0.2.

3.2 Experimental Results

The left two rows of Table 3 shows the composition of an ACTNET which was learned in the experiment 1 using a small data set. The number of classes was automatically determined by the class division in the I-PLCA. Fig. 1 shows a part of this ACTNET and the detail of an activity network of this ACTNET is shown in Fig. 4. The left row of Table 4 shows results of recognition and inference for action sequences used for learning. The classification accuracy of activities

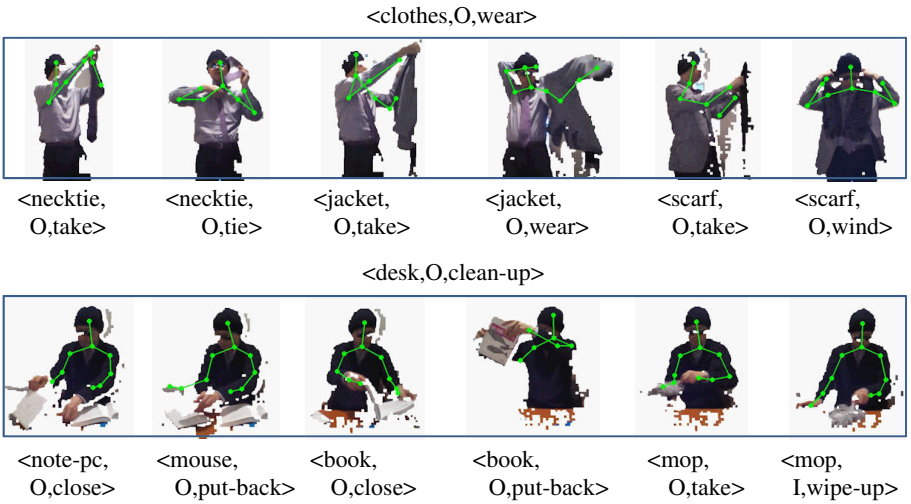

Fig. 3. Examples of action sequences in activities

Table 3. The composition of ACTNETs

Experiments	Experiment 1		Experiment 2	
	Activity	Action	Activity	Action
The number of classes	2	16	12	53.5
The number of target synsets	2	5	4	10
The number of motion synsets	2	10	4	16
The number of pairs of target and motion synsets	2	16	4	27

was 100%. The classification accuracy of actions was 81.3% when activities were used as context, whereas it was 75.0% when activities were not used as context. When additional information about object labels was given, the classification accuracy of actions was increased up to 93.8%.

In the experiment 2 using a large data set, 4 video clips were prepared for each of 4 activities and recognition and inference were evaluated for action sequences through 4-fold cross validation. The right two rows of Table 3 shows the composition of ACTNETs and the right row of Table 4 shows results of recognition and inference of actions and activities. The classification accuracy of activities was 93.8%. The classification accuracy of actions was 62.5% when activities were used as context, whereas it was 53.3% when activities were not used as context. When additional information about object labels was given, the classification accuracy of actions without and with contextual activities was respectively increased up to 75.8% and 83.4%. The percentage values in parentheses in Table 4 are the classification accuracy of the top two guesses of actions and activities.

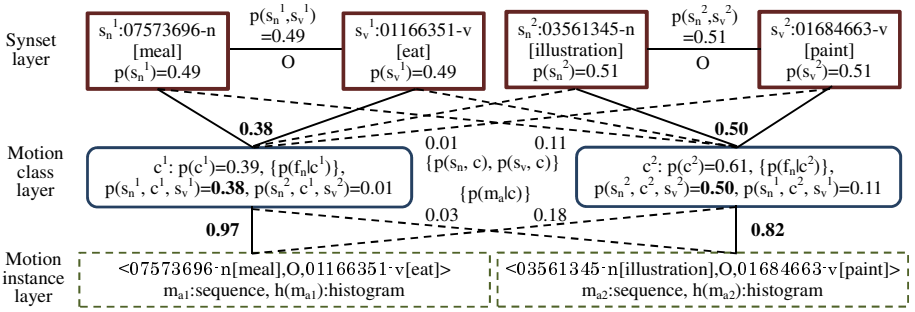

Fig. 4. An activity network of an ACTNET

Table 4. Results of recognition and inference

Experiments	Experiment 1	Experiment 2
Classification accuracy of activities	100%	93.8% (100%)
Classification accuracy of actions without contextual activities	75.0% (93.8%)	53.3% (59.2%)
Classification accuracy of actions with contextual activities	81.3% (100%)	62.5% (76.7%)
Classification accuracy of actions without contextual activities (when object labels are given)	93.8% (100%)	75.8% (85.8%)
Classification accuracy of actions with contextual activities (when object labels are given)	93.8% (100%)	83.4% (96.7%)

4 Discussion and Concluding Remarks

In data sets of experiments, there are the same actions for different objects and also different actions with similar motion, and these make it difficult to recognize object-oriented actions. In the experiment 2, classification accuracy is mainly decreased due to false recognition of 5 actions <notebook-computer,O, close>, <mouse,O,put-back>, <book,O,close>, <mop,O,take> and <teacup,O, pick-up>, which suffer the above difficulty and also have insufficient motion features since these actions are performed in a short time on a desk and skeleton extraction of the Kinect sensor is not accurate against the crowded background on the desk. In addition, in the data set of experiment 2, there are different activities which contains the same actions and just one false recognition of activities was occurred on one of these activities. As referential evaluation, though experimental setting and data sets are different, accuracy of similar action recognition by existing methods [7,8] is 48% $\sim$ 59% and our method achieved better results.

In conclusion, we have proposed a learning method of a probabilistic semantic network ACTNET which integrates visual features and semantic features of object-oriented actions and their contextual activities and also provided a method using the ACTNET which visually recognize and verbally infer actions

in the context of activities. Through two experiments, it has illustrated that the ACTNET can learn integrated probabilistic structure of visual and semantic features of object-oriented actions and activities and it has shown that activities and actions can be well recognized and inferred using the ACTNET, especially action understanding is improved using the context of activities and also additional information about objects.

Acknowledgment. This work was supported in part by Grant-in-Aid for Scientific Research (C) No.23500188 from Japan Society for Promotion of Science.

References

1. Bar, M.: Visual objects in context. Nature Reviews Neuroscience 5, 617–629 (2004)
2. Rabinovich, A., Vedaldi, C., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: Proc. of IEEE Int. Conf. on Computer Vision (2007)
3. Atsumi, M.: Object categorization in context based on probabilistic learning of classification tree with boosted features and co-occurrence structure. In: Bebis, G., et al. (eds.) ISVC 2013, Part I. LNCS, vol. 8033, pp. 416–426. Springer, Heidelberg (2013)
4. Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K.: Development of Japanese WordNet. In: Proc. of the 6th Int. Conf. on Language Resources and Evaluation, pp. 2420–2423 (2008)
5. Atsumi, M.: Learning visual categories based on probabilistic latent component models with semi-supervised labeling. GSTF Int. Journal on Computing 2(1), 88–93 (2012)
6. Kitani, K., Okabe, T., Sato, Y.: Discovering primitive action categories by leveraging relevant visual context. In: Proc. of the IEEE Int. WS on Visual Surveillance (2008)
7. Yao, B., Fei-Fei, L.: Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE Trans. on Pattern Analysis and Machine Intelligence 34(9), 1691–1703 (2012)
8. Yao, B., Ma, J., Fei-Fei, L.: Discovering object functionality. In: Proc. of Int. Conf. on Computer Vision (2013)