Semantic Action Recognition by Learning a Pose Lexicon

Lijuan Zhou, Wanqing Li, Philip Ogunbona, Zhengyou Zhang

PII: S0031-3203(17)30259-5
DOI: 10.1016/j.patcog.2017.06.035
Reference: PR 6200

To appear in: *Pattern Recognition*

Received date: 31 January 2017
Revised date: 11 June 2017
Accepted date: 30 June 2017

Please cite this article as: Lijuan Zhou, Wanqing Li, Philip Ogunbona, Zhengyou Zhang, Semantic Action Recognition by Learning a Pose Lexicon, *Pattern Recognition* (2017), doi: 10.1016/j.patcog.2017.06.035

Hihglihgts

- A novel semantic representation, pose lexicon, is proposed for action recognition.
- An extended hidden Markov alignment model is developed to learn a pose lexicon.
- Develop a semantic recognition method that is capable of zero-shot recognition.
- The proposed learning and recognition algorithms were evaluated on five datasets.

Semantic Action Recognition by Learning a Pose Lexicon

Lijuan Zhou[a], Wanqing Li[a,*], Philip Ogunbona[a], Zhengyou Zhang[b]

[a]*School of Computing and Information Technology, University of Wollongong, NSW 2522, Australia*
[b]*Microsoft Research, Redmond, WA 98052, USA*

Abstract

This paper proposes a semantic representation, *pose lexicon*, for action recognition. The lexicon is composed of a set of semantic poses, a set of visual poses and a probabilistic mapping between the visual and semantic poses. Specially, an action can be represented by a sequence of semantic poses extracted from an associated textual instruction. Visual frames of the action are considered to be generated from a sequence of hidden visual poses. To learn the lexicon, a visual pose model is learned from training samples by a Gaussian Mixture model to characterize the likelihood of an observed visual frame being generated by a visual pose. A pose lexicon model is also learned by an extended hidden Markov alignment model to encode the probabilistic mapping between hidden visual poses and semantic poses sequences. With the lexicon, action classification is formulated as a problem of finding the maximum posterior probability of a given sequence of visual frames that fits to a given sequence of semantic poses through the most likely visual pose and alignment sequences. The efficacy of the proposed method was evaluated on MSRC-12, WorkoutSU-10, WorkoutUOW-18, Combined-15 and Combined-17 action datasets using cross-subject, cross-dataset and zero-shot protocols.

Keywords: Lexicon, semantic pose, visual pose, action recognition

1. Introduction

Human action recognition is one of the most active research topics in computer vision due to its wide range of applications in smart video surveillance, human computer interaction, robotics, health care, etc. Over the past decade, different approaches have been proposed for recognizing human actions either from RGB video, depth map or skeleton data by associating low or middle level visual spatio-temporal features directly with class labels [1, 2, 3, 4, 5]. Specifically, many methods [1, 6, 7, 8] are based on the concept that an action can be well represented by a sequence of key or salient poses and these salient poses can be identified through visual features alone. However, these salient poses often do not

*Corresponding author.
Email addresses: lz683@uowmail.edu.au (Lijuan Zhou), wanqing@uow.edu.au (Wanqing Li), philipo@uow.edu.au (Philip Ogunbona), zhang@microsoft.com (Zhengyou Zhang)

necessarily possess semantic significance thus leading to the so-called semantic gap. We refer to the salient poses defined using visual features as *visual poses*.

On the other hand, how an action is or should be performed can be described by textual instructions. Such instruction usually follow Talmy's typology of motion [9] including four semantic elements: *Motion* (M), *Figure* (F), *Ground* (G) and *Path* (P); where M refers to the manner or direction of the movement, F represents the entity that changes its location following by a path P with respect to the reference object G. Noting that the combination of F, P, G and adverb or adjective phrases actually construct a linguistic description of the statuses or configurations of body parts that have to be gone through if the action is performed as instructed. These statuses or configurations defined in the linguistic description are referred to as *semantic poses*. Hence, an action can be described by a sequence of semantic poses if *F* is defined as the whole human body. Alternatively, if *F* refers to a body part, multiple sequences of semantic poses can be used. This observation leads to an opportunity for learning visual poses with semantic significance. Specifically, the sementic poses can be used to guide the learning of visual poses and visual poses can be probabilistically assigned with semantic meaning through their associations with semantic poses.

This paper proposes to learn a *pose lexicon* for semantic action recognition. The lexicon is composed of a set of semantic poses, a set of visual poses and a probabilistic mapping between the visual and semantic poses. A novel recognition method is developed upon the lexicon. To learn the lexicon, it is assumed that for an action there is an associated textual instruction available in training from which a sequence of semantic poses can be extracted through natural language parsing [10]. The visual frames of an action sample are considered to be generated from a sequence of hidden visual poses. To construct the lexicon, the key problems are to learn (1) a *visual pose model* from training samples that characterize the likelihood of an observed visual frame being generated by a visual pose and (2) a probabilistic mapping, referred to as a *pose lexicon model*, between the hidden visual poses and the semantic poses. In this paper, visual poses are modelled by a Gaussian Mixture Model and the mapping is represented by an alignment sequence which probabilistically associates visual poses with semantic poses. The alignment is hidden as the information is not available in either training or testing and is modelled by a hidden Markov alignment model. Both the visual pose model and pose lexicon model are learned using the Expectation-Maximisation strategy. Note that both visual poses and semantic poses are shared among all actions, one visual pose can be associated with one semantic pose and one semantic pose may be associated with multiple visual poses. With the lexicon, action classification is formulated as a problem of finding the maximum posterior probability of a given sequence of visual frames that fits to a given sequence of semantic poses

through the most likely visual pose and alignment sequences, which is solved by a modified Viterbi decoding algorithm in this paper.

The proposed lexicon and action recognition method have several advantages over the previous action recognition methods. First, opposed to the methods based on visual poses [11] or action topic models [12, 13], the proposed method bridges the semantic gap between visual frames and semantic (i.e. textual) description and offers a potential approach to semantic summarization of human motion and synthesis of motion based on textual description. Second, temporal characteristics of actions is captured by representing actions as either a sequence of semantic poses or a sequence of visual poses, however, it is ignored when representing actions by a set of attributes [14, 15]. Third, because the lexicon is shared by all actions, learning the visual pose model and pose lexicon model requires much less training samples than the cases where each action has its own lexicon or individual hidden Markov model is trained on each action as [16]. Fourth, the lexicon is automatically learned for encoding the relationship between visual and semantic poses. Such learning process is not given in the methods using activity triple ⟨subject, verb, object⟩ [17] or action bank [18] as semantic representation. Fifth, learning the lexicon does not require the mapping between visual frames and semantic poses sequences, which reduces the cost of manual annotation as required by the method [19]. Last, actions that have not been specifically included in the training can be recognized without expanding the pose lexicon model as learned by action graph [1].

This paper is an extension of our previous work [20]. The following are the additional contributions of this work compared to the earlier version.

(1) For computational simplicity, salient frames are extracted rather than using all frames of action video. Two new methods of extracting salient frames are proposed and evaluated. In [20], salient frames are assumed to be frames in which human body reaches maximum or minimum extensions. Here, this assumption is extended to include frames in which motion reaches maximum or minimum speed.

(2) Unlike a separate process adopted in [20], we jointly generate visual pose sequences and align them to semantic pose sequence.

(3) Temporal constraint is incorporated to the pose lexicon model for capturing temporal characteristic of actions.

(4) The proposed model is evaluated on three more datasets: WorkoutUOW-18, Combined-15 and Combined-17. Comparisons are included with skeleton-based recognition methods, methods based on semantic learning. In addition, cross-dataset and zero-shot recognition were performed to verify the transferring knowledge of pose lexicon.

The rest of this paper is organized as follows. Section 2 provides a review of previous work related to semantic action recognition, alignment of video with language and language parsing. The proposed model is formulated and solved in Section 3. Section 4 presents an recognition system based on skeleton data. In Section 5, experiments are presented to demonstrate the efficacy of the proposed method. Finally, the paper is concluded with remarks in Section 6.

2. Related work

This section provides an overview of literature in the areas related to our work, including semantic action recognition, language assisted video analysis and extraction of semantics from language.

Semantic action recognition: Despite the progress made in action recognition over the past decade, few studies have been reported based on semantic representation. Earlier methods [1, 11, 21] represent actions by visual poses and the visual poses are learned from visual features only. Later, actions are represented by action topics [12, 22, 23] using topic models such as probabilistic latent semantic analysis [24] and latent Dirichlet allocation (LDA) [25]. Here, action topics are learned from visual pose where visual poses are considered as action words. The intuitive basis of using visual poses and action topics is that frequently co-occurring low-level features are correlated to some degree with some conceptual entities with high-level features. It is noteworthy that both methods are bottom-up approach without guaranteeing that learned features have semantics. However, the semantic poses extracted from textural instructions offer such warranty.

Apart from learning latent action semantics by topic models, another approach focused on defining explicit semantic elements to describe action or activity related properties. The first kind of semantic elements are attributes which were introduced by Liu et al. [14] and further improved by Zhang et al. [26]. The attributes are descriptive language of action spatial characteristics such as "arm pendulum like motion" and "torso twist motion". Rohrbach et al [15] proposed to represent activities by attributes for the recognition of cooking activities only and the attributes are defined as basic-level activities and objects touched by hands. Action or activity descriptors are often built as histograms of attributes so that temporal characteristics of actions is hardly modelled by attributes. The second kind of semantic concepts are components that constitute activities or actions. Sadanand et al. [18] proposed to describe activities by action bank which consists of a set of action detectors. Each action detector includes the templates in different viewpoint and activities are represented as a histogram of action templates for training the classifier. Noting that action detectors encode the relationship between visual representation and semantic representation (i.e. action labels) and they can be considered as lexicon of our paper. However, action detectors were manually selected by authors of [18] which cannot be directly applied in the recognition

of other actions. Semantic hierarchy was built in [19, 27] for describing components that constitute both actions and activities. In [27], an activity is represented by three semantic levels: video sequence, action labels and activity label. In our paper, an action is represented by three levels: video sequence, visual pose sequence and semantic pose sequence, where semantic pose sequence is similar to the level of action label in [27]. We can see that [27] directly associates visual feature with class labels, while our paper uses visual pose sequence as a bridge to link visual features with semantic poses. The advantage of introducing the level of visual poses is that the trained model does not need to be re-trained for classification when new action or activity is added to the model. Like this paper, the paper [19] does not directly associate visual features with action labels in the semantic hierarchy. In particular, an activity in [19] is represented by four levels: video sequence, poses, actions and activity. The pose level is represented by a histogram of visual poses and the histogram was constructed by assigning a visual pose for each frame. Although latent assignments between visual frames and visual poses was applied in [19] same as our paper, the latent assignments were binary values (so-called as hard mapping) while this paper applies soft mapping. The action level is represented by a histogram over visual poses. Here, the assignments between video frames with action labels needs to be annotated before training, which is different with our problem because the assignment of video frames with semantic poses is unknown. The third kind of semantic elements are defined from the view of sentence components. Guadarrama et al. [17] proposed a method of generating a sentence for an input video. Based on the target of language generation, activity videos are described by subject, verb and object because the three components are actually main parts of a sentence. The triplet $\langle$subject, verb, object$\rangle$ interprets the actor, the action and the object manipulated by the actor and it can be understood as semantic contents of the activity video. The relationship between visual representation with action and objects is respectively encoded by a pre-learned action codebook and object detector, thus, it is available during both training and test. However, such relationship is encoded in the pose lexicon needs to be learned in our paper.

Linguistics assisted video analysis: The main tasks of linguistics assisted video analysis include action recognition, natural language generation of videos, alignment of video with text. With the assistance of language, action recognition can be applied based on learning semantic contents [17] and traditional action recognition using visual features can be improved through mining semantic relationship of action and object names[28].

The task of natural language generation for videos involves to answer the following questions: "What is worth describing in the video? Which words are suitable for describing the video? How to generate a smooth sentence?" The methods for this task can be divided into two kinds: (1) use a two stage pipeline that first identifies the semantic contents

(i.e. subject, verb, object) and then generate a sentence based on a template [17, 29]; (2) avoid the separation of content identification and sentence generation by learning the probability distribution in the common space of visual content and textual sentence [30, 31, 32].

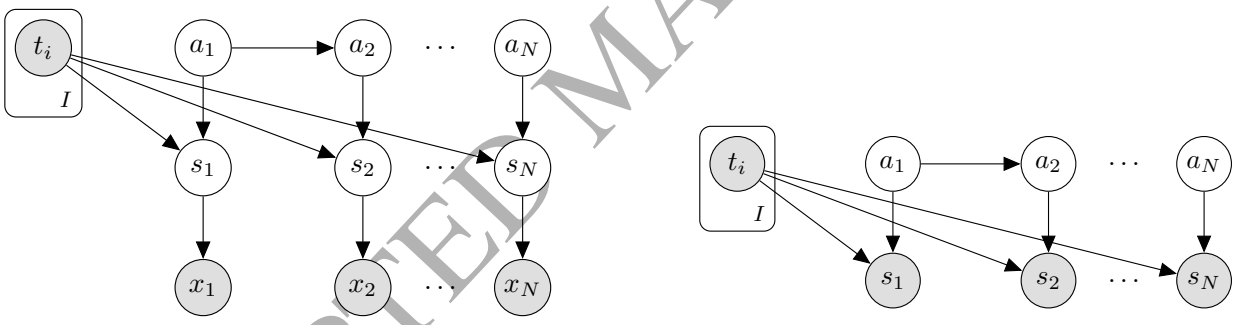
The task of alignment of video with text aims to map videos with corresponding sentences. Given a sequence of semantic poses and a sequence of video frames, we would like to learn the mapping of semantic pose with their referents in the videos. The pose lexicon model belongs to the problem of aligning video with text. Since manually acquiring the alignment can be tedious and expensive for large collection of parallel video and text datasets, unsupervised alignment is crucial for scaling to large datasets. Naim et al. [33] introduced a generative model for unsupervised alignment of protocol sentences with video segments. Each video frame is segmented into a set of blobs. Two-level hierarchical alignments were performed for simultaneously aligning protocol sentences to corresponding video frames and matching nouns in the sentences to corresponding blobs in the video frame. A hidden Markov alignment model [34] was learned for the higher-level alignment and IBM model 1 [35] was trained for lower-level alignment. We extend the hidden Markov alignment model by incorporating hidden variables between hidden states and observations. Later, for the same alignment problem Naim et al. [36] also proposed a discriminative unsupervised model based on latent variable conditional random filed [37]. Bojanowski et al. [38] presented a discriminative clustering approach using an inter quadratic program to learn an implicit linear mapping between low-level action feature and text feature, whereas our model learns a mapping between mid-level feature (visual pose) and text. Such generalization of low-level visual representation is beneficial for motion generation. Recently, Song et al. [39] aligned action hyper-features to verbs in the corresponding sentence, which involves two separate steps: generate action hyper-features by accumulating low-level visual features based on a bag-of-feature approach; match hyper-features with verbs. The hyper-feature are conceptually similar to features extracted by deep models but without supervision, thus, they are still visual features. The essential difference between the video-text alignment task and our pose lexicon model is the former often directly map visual features to text and the latter introduces the visual pose which can replace same visual feature (eg. bag-of-words feature used in the video-text alignment task). There is no obvious advantage of using visual pose or low-level visual feature in the alignment task but there are advantages in action classification problem.

Extracting semantics from language: Semantics can be extracted from textual instructions using syntactic parsing. Currently, there are two types of parsers including constituency- and dependence-based parsers. Constituency-based parser decomposes a sentence into phrases, such as noun phrases, verb phrases, preposition phrases, which focuses on

constituency relation between phrases. Such parsers are seen from Charniak parsers [40], Stanford parser [41] and Berkeley parser [10]. Dependence-based parser focuses on dependent relation between words, for example, Stanford [42], MST [43], Bohnet's [44] and Malt parsers [45]. For the purpose of extracting the semantic poses from instructions, constituency-based parser is sufficient for generating noun, verb and preposition phrases and matching to the elements of *Motion*, *Figure*, *Ground* and *Path*. Details are beyond the scope of the paper.

3. Proposed method

Let $X = \{x_n\}_{n=1}^{N}$ be a sequence of visual frames representing an action and x_n be the feature vector extracted from frame n. The corresponding semantic pose sequence is denoted as $T = \{t_i\}_{i=1}^{I}$ ($t_i \in \Psi$), where $\Psi = \{\psi_g\}_{g=1}^{G}$ is the set of semantic poses. The hidden visual pose sequence of this action is written as $S = \{s_n\}_{n=1}^{N}$ ($s_n \in \Omega$), where $\Omega = \{\omega_h\}_{h=1}^{H}$ is the set of visual poses. Figure 1a is the graphical model illustrating the relationship among X, T, S and alignment A, where the non-shaded nodes indicate hidden variables and its comparison to the conventional hidden Markov alignment model as shown in Figure 1b.



(a) The proposed model (b) Conventional hidden Markov alignment model [34]

Figure 1: The proposed model and its comparison to the conventional hidden Markov alignment model. In the figure, shaded circles represent observable variables and non-shaded circles are hidden variables. The goal of the proposed model is to learn both visual pose sequence $S = \{s_1, \ldots, s_N\}$ and alignment sequence $A = \{a_1, \ldots, a_N\}$ while the conventional hidden Markov model only needs to learn the alignment sequence.

Let $\mathcal{Y} = \{T_r\}_{r=1}^{R}$ be a set of actions, where T_r is the semantic pose sequence of action r and each element of T_r

belongs to Ψ. The problem of action classification is to find the semantic pose sequence that most likely explains X.

$$\begin{aligned} T^* &= \arg\max_{T\in\mathcal{Y}} P(T|X) \\ &= \arg\max_{T\in\mathcal{Y}} \frac{P(T)P(X|T)}{P(X)} \\ &\propto \arg\max_{T\in\mathcal{Y}} P(T)P(X|T), \end{aligned} \tag{1}$$

where $P(T)$ is prior probability of the semantic pose sequence T and $P(X|T)$ is the likelihood of the semantic pose sequence explains the visual feature sequence. $P(T)$ depends on the application and in this paper it is assumed to be uniform across the actions in $\mathcal{Y}$. Hence, the problem becomes to maximise the likelihood $P(X|T)$.

As discussed above, the visual feature sequence, X, is considered to be generated from a hidden visual pose sequence, S, whose elements belong to the visual pose set Ω. Therefore,

$$P(X|T) = \sum_S P(X, S|T) = \sum_S P(S|T)P(X|S, T). \tag{2}$$

where $P(S|T)$ in Eq.(2) represents the mapping between a visual pose sequence and a semantic pose sequence and $P(X|S, T)$ represents the probability of X being generated from S given T.

To simplify the calculation of $P(X|S, T)$, it is assumed that X and T are conditionally independent of each other given S and that each frame of the visual feature sequence is generated independently from its corresponding visual pose. Hence,

$$P(X|S, T) = P(X|S) = \prod_{n=1}^{N} P(x_n|s_n). \tag{3}$$

$P(x_n|s_n)$ is the likelihood of generating video frame x_n by visual pose s_n, referred to as a *visual pose model*, where $s_n \in \Omega$.

$P(S|T)$ measures the probability of associating sequence T of semantic poses to sequence S of visual poses. This problem has an analogy to the widely studied word alignment problem [34, 46] in machine translation if T and S were treated as word sequences of two languages. Specifically, one semantic pose can be associated with multiple visual poses and one visual pose can only be associated with one but only one semantic pose. However, unlike word alignment in translation, the alignment of the elements in S and those in T has strict temporal constraint. In addition, element s_n in S can be any of the poses Ω with a probability, rather than a fixed word in the translation.

Let $A = a_1^N = \{a_1, ..., a_n, ..., a_N\}$ be an alignment vector, where $a_n \in [1, I]$ refers to as the aligned position of n-th visual pose in the semantic pose sequence, N and I are respectively the length of visual pose sequence and length of the

semantic pose sequence. For example, if the n^{th} visual pose is associated with i^{th} semantic pose, $a_n = i$. Based on the Markov assumption, $P(S|T)$ can be expressed as

$$P(S|T) = \sum_{A} P(S, A|T) = \sum_{a_1,\ldots,a_N} P(a_1)P(s_1|t_{a_1}) \prod_{n=2}^{N} P(a_n|a_{n-1})P(s_n|t_{a_n}), \tag{4}$$

where $P(s_n|t_{a_n})$ is the probability of associating $s_n, s_n \in \Omega$ with $t_{a_n}, t_{a_n} \in \Psi$ and $P(a_n|a_{n-1})$ is transition probability. $P(a_1)$ is the probability of aligning the first visual pose. The many-to-one association between visual poses and semantic poses yields

$$\sum_{\omega_h \in \Omega} P(\omega_h|\psi_g) = 1. \tag{5}$$

$P(\omega_h|\psi_g)$, referred to as a *pose lexicon model*, is the lexical probability of associating $\psi_g \in \Psi$ with $\omega_h \in \Omega$.

Unlike word alignment in translation where a word in one language can be aligned with any words in the other language, the association of a visual pose with a semantic pose has to follow a strict temporal constraint. In particular, a visual pose can only be associated with the same or next semantic pose in the sequence that the previous visual pose is associated with. Let a_n and a_{n-1} be the associated position of n-th visual pose and its previous visual pose. Therefore, a_n can only be a_{n-1} or $a_{n-1} + 1$, where the former means that the n^{th} visual pose is associated with the same semantic pose of the $(n-1)^{th}$ one and the latter means that the n^{th} visual pose is associated with the next semantic pose of the $(n-1)^{th}$ one. This constraint does not affect the alignment of actions with repetitive semantic poses (eg. "walk", "jog") because the alignment depends on multiply score of transition and lexical probability. Let P_0 is the probability of being $a_n = a_{n-1}$, i.e. how likely a visual pose is associated with the same semantic pose. The temporal constraint can the be explicitly expressed as

$$P(a_n|a_{n-1}) = \begin{cases} P_0 & a_n - a_{n-1} = 0 \\ 1 - P_0 & a_n - a_{n-1} = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

In addition, the first visual pose is assumed to be associated with the first semantic pose and each semantic pose has at least one associating visual pose.

$$P(a_1) = \begin{cases} 1 & a_1 = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Substitute Eqs.(3) and (4) for Eq.(2), $P(X|T)$, the probability that a sequence X of a video frames can be explained

by a sequence T of semantic poses extracted from textural instructions, is

$$P(X|T) = \sum_{s_1,\ldots,s_N} \sum_{a_1,\ldots,a_N} P(a_1)P(s_1|t_{a_1})P(x_1|s_1) \prod_{n=2}^{N} P(a_n|a_{n-1})P(s_n|t_{a_n})P(x_n|s_n). \quad (8)$$

3.1. Learning

To classify a visual sequence using Eq.(8), the visual pose model $P(x_n|s_n)$, $s_n \in \Omega$, pose lexicon model $P(\omega_h|\psi_g)$, $\omega_h \in \Omega$ and $\psi_g \in \Psi$, and parameter P_0 need to be learned from training samples. Joint learning of these models are possible, but the computational cost would be high. A common low-cost approach is to parametrize the visual pose model and learn it separately from the pose lexicon model through unsupervised clustering, for instance a Gaussian Mixture Model.

Given $P(x_n|s_n)$, learning P_0 and $P(\omega_h|\psi_g)$ needs to maximize Eq.(8) from the exponential number (in precise, $(IH)^N$) of possible combinations of visual pose sequence S and the alignment A, which makes maximization of Eq.(8) is intractable. In [20], the computation is further simplified by first determining the visual pose sequence by assigning the most likely visual pose ω_k to a visual frame x_n based on the visual pose model $P(x_n|\omega_k) \geq P(x_n|s_n)$, $s_n \in \Omega$ and $s_n \neq \omega_k$. Then, the problem is reduced to finding the alignment A given S and T. More details can be found in [20]. The disadvantage of this method is that the quality of the learned pose lexicon model depends much on the quality of the visual pose sequence, hence, is sensitive to the noise in the visual frames.

To improve the robustness to noise, the visual pose sequence S in this paper is determined simultaneously with the alignment sequence A. Expectation-maximization (EM) strategy is employed to jointly estimate parameters $\theta = \{P(\omega_h|\psi_g), P_0\}$ and hidden variables S and A. The EM algorithm finds the maximum marginal likelihood and alternately conducts the following two steps.

E-step: Posterior probabilities among alignments and visual poses are calculated for estimating the hidden visual pose sequence and alignment. Since the full model is a hidden Markov model, the forward-backward algorithm is employed to reduce the computation complexity.

Let forward probability $\alpha_i(n)$ be the joint probability of observing video frame sequence $\{x_1, \ldots, x_n\}$ and associating n-th visual pose to the i^{th} semantic pose and the forward recursion is given as follows

$$\alpha_i(n) = \Big[P_0\alpha_i(n-1) + (1-P_0)\alpha_{i-1}(n-1)(1-\delta(i-1))\Big] \sum_{s_n \in \Omega} P(s_n|t_i)P(x_n|s_n), \quad (9)$$

where $\delta(\cdot)$ is a Kronecker delta function. The joint probability of observing visual frame sequence $\{x_1, \ldots, x_n\}$ and

associating n^{th} visual pose to i^{th} semantic pose through $s_n = \omega_h$, denoted as $\alpha_i^h(n)$, can be expressed as

$$\alpha_i^h(n) = \Big[P_0\alpha_i(n-1) + (1-P_0)\alpha_{i-1}(n-1)(1-\delta(i-1))\Big]P(s_n = \omega_h|t_i)P(x_n|s_n = \omega_h), \tag{10}$$

The backward probability $\beta_i(n)$ measures the probability of observing the visual frame sequence $\{x_{n+1}, \ldots, x_N\}$ when n^{th} visual pose is associated with i^{th} semantic pose and can be expressed recursively as

$$\beta_i(n) = \sum_{s_{n+1}\in\Omega} P(x_{n+1}|s_{n+1})\Big[P_0\beta_i(n+1)P(s_{n+1}|t_i) + (1-P_0)\beta_{i+1}(n+1)P(s_{n+1}|t_{i+1})(1-\delta(i-I))\Big]. \tag{11}$$

Let backward probability $\beta_i^h(n)$ be the conditional probability of observing visual frame sequence $\{x_{n+1}, \ldots, x_N\}$ given the knowledge that n^{th} visual pose is generated from i^{th} semantic pose and $s_n = \omega_h$. Since $\beta_i(n)$ is independent with s_n, $\beta_i^h(n) = \beta_i(n)$.

With the forward and backward probabilities, the probability of associating n^{th} visual pose with the i^{th} semantic pose and inferring $s_n = \omega_h$ is defined as follows

$$\gamma_i^h(n) = P(a_n = i, s_n = \omega_h|X, \theta) = \frac{\alpha_i^h(n)\beta_i^h(n)}{\sum_{i=1}^I \alpha_i^h(n)\beta_i^h(n)}. \tag{12}$$

The joint probability that the $(n-1)^{th}$ and n^{th} visual poses are associated with the i^{th} semantic pose is

$$\xi_{i,i}(n-1) = P(a_{n-1} = i, a_n = i|X, \theta) = \frac{1}{Z}P_0\alpha_i(n-1)P(s_n|t_i)\beta_i(n). \tag{13}$$

Similarly, the joint probability of associating $(n-1)^{th}$ and n^{th} visual poses respectively from i^{th} and $(i+1)^{th}$ semantic poses is calculated by

$$\xi_{i,i+1}(n-1) = P(a_{n-1} = i, a_n = i+1|X, \theta) = \frac{1}{Z}(1-P_0)\alpha_i(n-1)P(s_n|t_{i+1})\beta_{i+1}(n), \tag{14}$$

where

$$Z = \sum_{i=1}^{I}\alpha_i(n-1)\Big[P_0P(s_n|t_i)\beta_i(n) + (1-P_0)P(s_n|t_{i+1})\beta_{i+1}(n)(1-\delta(i-I))\Big].$$

M-step: Parameters θ is re-estimated by maximizing the posterior probability. In particular, $P(\omega_h|\psi_g)$ is re-estimated through

$$P(\omega_h|\psi_g) = \frac{\sum_{\{T,X\}\in D}\sum_{n=1}^N\sum_{i=1}^I \gamma_i^h(n)\delta(t_i - \psi_g)}{\sum_{\omega_h}\sum_{\{T,X\}\in D}\sum_{n=1}^N\sum_{i=1}^I \gamma_i^h(n)\delta(t_i - \psi_g)}. \tag{15}$$

The staying probability P_0 is updated through

$$P_0 = \frac{\sum_{\{T,X\}\in D}\sum_{n=2}^N\sum_{i=1}^I \xi_{i,i}(n-1)}{\sum_{\{T,X\}\in D}\sum_{n=2}^N\sum_{i=1}^I \Big[\xi_{i,i}(n-1) + \xi_{i,i+1}(n-1)(1-\delta(i-I))\Big]}, \tag{16}$$

where D represents the set of visual frame and semantic pose sequences of all training samples.

3.2. Classification

Given a visual sequence X, the problem of action classification is to find the most likely semantic pose sequence that explains X. It is performed in two steps: (1) for any action $T \in \mathcal{Y}$ find the most likely alignment sequence and visual pose sequence that maximize $P(X|T)$ as expressed by Eq.(1), (2) classify X as action T_r that $P(X|T_r)$ is maximum.

Since the alignment sequence is a Markov chain, Viterbi decoding algorithm is employed to find the best alignment. Let $V_i(n)$ be the probability of the most likely alignment for the first n visual frames with the alignment of the n^{th} visual frame to the i^{th} semantic pose. Suppose the length of X is N, maximizing $P(X|T)$ equals to maximizing $V_i(N)$ which can be recursively calculated

$$V_i(n) = \max_{i' \in \{i, i-1\}} \Big\{ V_{i'}(n-1) P(a_n = i | a_{n-1} = i') \max_{s_n \in \Omega} \{ P(s_n|t_i) P(x_n|s_n) \} \Big\}. \tag{17}$$

Noting that the second term takes the likelihood of a frame x_n being generated from a visual pose into consideration. This is different from conventional decoding in word alignment where x_n would certainly rather than probably belong to one word.

The time complexity of Viterbi decoding for finding the best alignment in hidden Markov alignment model is $\mathcal{O}(NI^2)$. However, this is reduced to $\mathcal{O}(IN)$ because of the strict temporal constraint. Since the determination of visual poses uses the complexity of $\mathcal{O}(H)$, the total complexity of finding the most likely alignment and visual pose sequences is $\mathcal{O}(HIN)$.

4. A skeleton-based action recognition system

This section presents a system to recognize actions from skeleton data to validate the efficacy of the proposed method. Note that the the proposed method can be applied to other modalities like RGB video and depth maps in a similar way.

Figure 2 is the block-diagram of the system. In training, each action is assumed to have textual instructions from which its sequence of semantic poses can be parsed and to have multiple instances of skeletons sequences performed by multiple subjects. In testing, the system essentially determines how likely a testing sequence of skeletons is associated with a trained sequence of semantic poses or a semantic pose sequence of an action that is defined by textual instructions and has not been included in the training, but all of its semantic poses can be parsed from the available textural instructions and are seen in the learned lexicon. The former is referred to as *recognition of trained actions* and the latter is called *zero-shot recognition*. To reduce the computation, salient frames are extracted from each skeleton sequence before training and

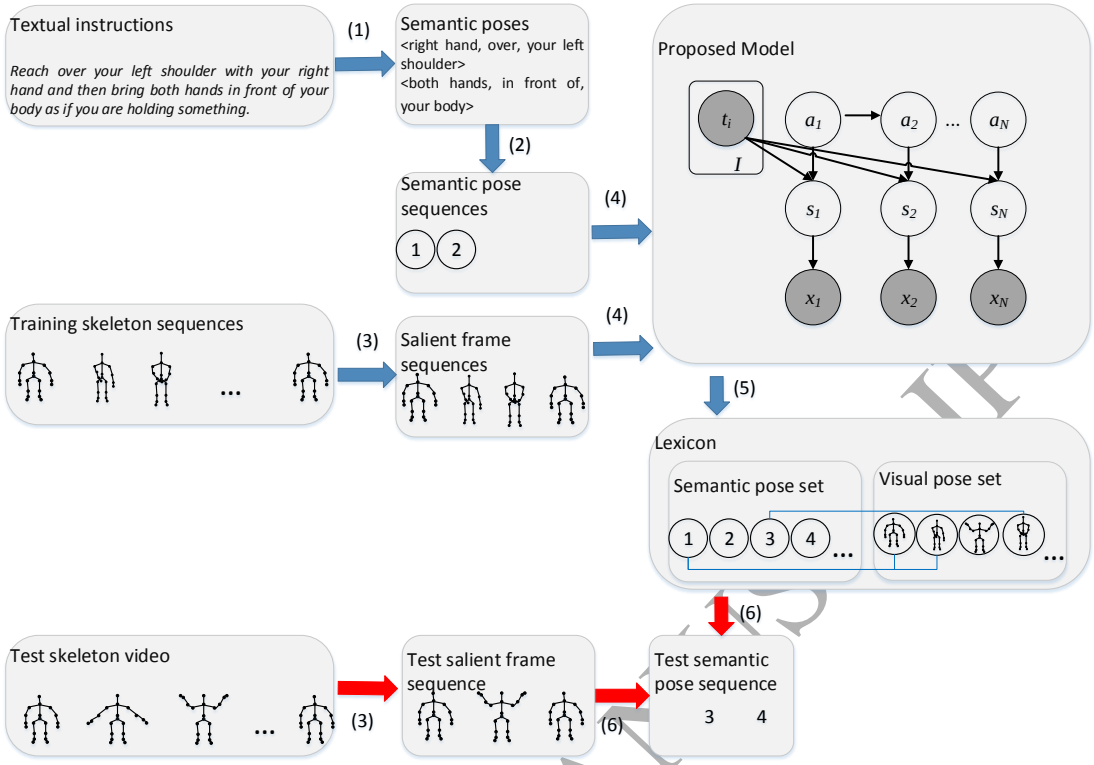


Figure 2: Framework of the proposed method. Training (blue arrows) follows steps (1-5) and test (red arrows) follows step (3) and (6). Here, step (1): extract semantic poses. Step (2): generate symbolized semantic pose sequences by clustering semantic poses to a semantic pose set. Step (3): extract salient frames and cluster all salient frames to generate a visual pose set. Step (4): train the proposed model. Step (5): obtain the learned pose lexicon. Step (6): assign test salient frame sequence to the class with the most likely semantic pose sequence.

testing to reduce the number of frames for processing. Those salient frames serve as the visual frames X in the proposed method.

4.1. Semantic poses extraction

Textual instructions of an action consist of by four semantic elements including *Motion* (M), *Figure* (F), *Ground* (G) and *Path* (P). Semantic poses are parsed from textural instructions based on the three semantic elements, F, G and P. The manner represented by M is ignored in this paper because the manner does not contribute semantically to the category of the motion and the motion direction is extracted from the path P though M may also have the information of motion direction. F is normally encoded by noun phrases near M (encoded by verb) while P and G are normally encoded by prepositional phrases in the textual instructions.

Note that parsing of textual instructions is not the focus of this paper. Nonetheless, a brief description is given in this

section. One approach to extracting the semantic poses from textual instructions is to use a Constituency-based parser (i.e. Berkeley [10]) to obtain three-tuples $< \mathrm{F}, \mathrm{P}, \mathrm{G} >$ where each three-tuples describes as a semantic pose. Since ground G often includes source, mediums and target, there may be multiple three-tuples arranged in a time order, so that a sequence of semantic poses are generated. A semantic pose set is generated by grouping same semantic poses used in different actions and symbolized semantic pose sequences can be produced accordingly. Take the textual instruction of the action "change weapon" shown in Section 1 as an example, the extracted three-tuples are "<right hand, over, your left shoulder>, <hands, in front of, your body>". Suppose the semantic pose set is written as $\{\psi_1, \psi_2\}$ where ψ_1 and ψ_2 respectively represents "<right hand, over, your left shoulder>" and "<hands, in front of, your body>". The symbolized semantic pose sequence of the example is written as $\{\psi_1, \psi_2\}$.

4.2. *Extraction of salient frames*

It is observed that semantic poses presented in textual instructions often correspond to the flexion and extension of body parts. Therefore, salient frames are defined in this paper as those at which one or multiple body parts reach their minimum/maximum spatial positions or have minimum/maximum motion or change of motion. Specifically, salient frames are determined based on the concepts of potential and kinetic energy in mechanics. Given an action instance $O = \{o_{fj}\}_{f=1:F}^{j=1:J}$ containing F frames, where $\{o_{fj}\}^{j=1:J}$ represents the normalized skeleton [47] at frame f with J joints and o_{fj} refers to $3D$ positions of the j-th joint.

Potential energy (PE) at frame f is calculated with respect to its neutral pose. A neutral pose is defined as the pose of the body standing upright with the two arms at rest position as shown in Fig. 3a.

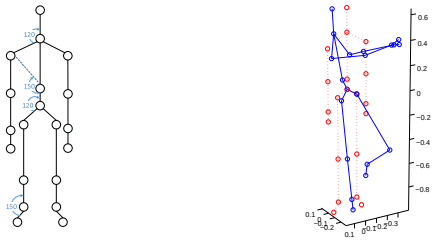
$$PE_f = \sum_{j=1}^{J} (o_{jf} - o_{jf}^c)^2, \tag{18}$$

where o_{jf}^c is the neutral position of j-th joint at frame f; which is considered as the reference position. PE captures how much extension of body parts. The difference between potential energy (PE) and eigen-decomposition (ED) of covariance in [20] is that PE considers extensions of all joints whereas ED only considers the major axis of the extensions.

Kinetic energy (KE) is the energy due to motion and is simply calculated as

$$KE_f = \sum_{j=1}^{J} (o_{j(f+1)} - o_{j(f-1)})^2. \tag{19}$$

For the action instance O, its PE sequence is represented by $\{PE_1, ..., PE_f, ..., PE_F\}$ and KE sequence is written as $\{KE_1, ..., KE_f, ..., KE_F\}$. The total energy (TE) sequences is obtained by $\{PE_1 + KE_1, ..., PE_f + KE_f, ..., PE_F +$



(a) Neutral pose (b) A frame and its neutral pose

Figure 3: Visualization of neutral pose.

$KE_F\}$. These sequences are filtered by a Gaussian filter with a window size of 3 frames to reduce the affect of noise in the skeleton data. Frames at which PE, KE or TE reaches their local minima or maxima are considered to be salient frames. Figure 4 shows the plots of PE and KE sequences before and after Gaussian filtering for action "Life outstretched arms". The extracted salient frames based on PE and KE are respectively marked by circles and triangles. Sample salient frames extracted by TE are shown in Figure 5, where the numbers represent the frame number in Figure 4.

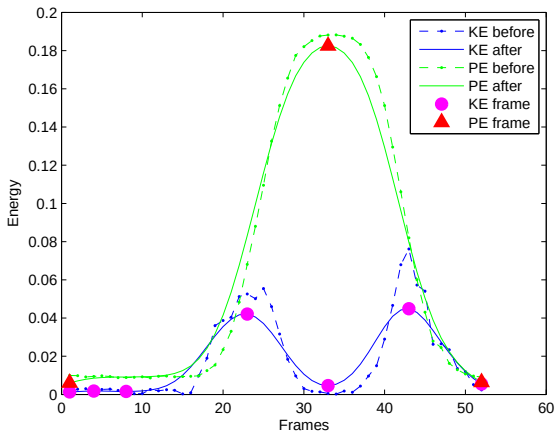


Figure 4: Illustration of extraction of salient frames based on potential and kinetic energy.

Figure 5: Visualization of sample salient frames extracted by TE.

4.3. Visual pose extraction

The salient frames of all action instances are fitted to a Gaussian Mixture Model (GMM) to generate a set of visual poses. In this paper, a moving pose descriptor [47] is extracted from each salient skeleton frame for the clustering. Each component of the GMM is considered as a visual pose and the number of components depends on the number of semantic poses. One semantic pose can be mapped to multiple visual poses when these visual poses are similar while one visual pose corresponds to at most one semantic pose. Therefore, the number of components is chosen to be larger than the number of semantic poses so that any semantic pose can correspond to at least one visual pose.

5. Experiments and results

To evaluate the efficacy of the proposed method, four experiments were conducted including: (1) salient frame extraction which evaluates how well the extracted frames represent the original action instances; (2) visual pose modelling for investigating the representation characteristics of the generated visual poses; (3) pose lexicon learning which evaluates the relationship between visual poses and semantic poses; (4) action classification for demonstrating the benefit of pose lexicon in action recognition.

Since one semantic pose relates to one or multiple visual poses, the number of semantic poses is considered as the indicator for setting the number H of visual poses. In the experiments (2)-(4), the value of H was set to 1-7 times as many as the number of semantic poses. Following the initialization of expectation maximization used in alignment models [35], the lexicon probability $P(\omega_h|\psi_g)$ and staying probability P_0 in the experiments (3)-(4) were initialized as uniform distribution. Therefore, $P(\omega_h|\psi_g)$ was set as $\frac{1}{H}$ for any $\omega_h \in \Omega$ and P_0 was set as 0.5.

5.1. Datasets

Five skeleton action datasets that cover a range of actions in human-computer actions and daily exercises were used to evaluate the proposed method: MSRC-12 Kinect gesture [48], WorkoutSU-10 exercise [49], WorkoutUOW-18 exercise, Combined-15 and Combined-17. The textual instructions of the actions in these datasets are provided in the supplemental material.

MSRC-12 Kinect gesture dataset [48]: This dataset was collected at Microsoft Research Cambridge. It comprises of 594 sequences of skeletal body part movements collected from 30 subjects performing 12 gestures. One sequence contains 9 to 11 instances. The gestures can be categorized into two abstract categories: Iconic and Metaphoric gestures.

The Iconic gestures imbue a correspondence between the gesture and the reference which include *duck, goggles, shoot, throw, change weapon, kick*. The Metaphoric gestures represent an abstract concept including *lift outstretched arms, push right, wind up, bow, had enough, beat*. The subjects were instructed through three types of modalities to perform actions. The instruction modalities are (1) descriptive text breaking down the performance kinematics; (2) an ordered series of static images; and (3) video (dynamic images). In the experiments of this dataset, same method as that in [50] was used to split the 594 sequences into 6244 instances for evaluation. Note that about half of the instances were recorded by informing the subject through descriptive text. There are 16 semantic poses in total extracted from the descriptive texts.

WorkoutSU-10 exercise dataset [49]: This dataset was collected at Sabanci University in Istanbul. It comprises of 1500 sequences in total, collected from 15 subjects performing 10 fitness exercises. Each exercise has been repeated 10 times by each subject. It contains more than 5 million frames and large time span within each instance. Exercise types of this dataset selected by professional coaches are grouped in three categories: Balance, Strength and Flexibility. The Balance exercises are *balance with hip flexion, balance-trunk rotation, lateral stepping*. The Strengthening exercises include *curl-to-press, freestanding squats, transverse horizontal dumbbell punch, oblique stretch*. The Flexibility exercises are *thoracic rotation bar on shoulder, hip adductor stretch (B2), hip adductor stretch (B3)*. The subjects were given descriptive texts and videos to perform exercises. Sixteen semantic poses were parsed from the descriptive texts.

WorkoutUOW-18 dataset: This dataset was newly created by the authors. It focuses on workout actions and consists of 18 exercises performed by 14 subjects. The exercises are *standing clamshells, balance, leg cycles, toe touch kick, high knee, prisoner squat, lunge with right kick, lunge with knee up, side to side with bar, squat with left kick, ninety raise, reverse curl and side press, standing L raise, sitting L raise, bent over row, bent over lateral raise, standing triceps extension and side to side bend*. The dataset was collected under the supervision of a professional trainer using a Microsoft Kinect v2 and has three modalities: RGB video, depth maps and skeletons. The subjects were given textual instructions and demonstration video on how the exercises should be performed. Considering subjects may remember how to perform the actions after watching video, subjects were required to perform the actions 3 times first based on the textual instruction and then they were asked to watch the demonstration video and perform the actions for 3 more times. There are 1509 valid action instances in total. The dataset was specifically designed for evaluating zero-shot recognition because there are many shared semantic poses among the actions. Totally, 41 semantic poses in total were parsed from the textural instructions of the actions in this dataset.

Combined-15 dataset: This dataset was constructed for evaluating cross-dataset action classification. It combines

Action	Source dataset 1	Source dataset 2	Action	Source dataset 1	Source dataset 2
bent	MSRC-12	UOW-Action 3D	cross arms	MAD	UTD-MHAD
right leg front kick	UOW-Action 3D	MAD	basketball shoot	MAD	UTD-MHAD
right leg side kick	UOW-Action 3D	MAD	clap hands	UTD-MHAD	UT-Kinect
throw	MAD	UTD-MHAD	stand to sit	UTD-MHAD	UT-Kinect
jog	MAD	UTD-MHAD	sit to stand	UTD-MHAD	UT-Kinect
right arm swipe left	MAD	UTD-MHAD	squat	WorkouSU-10	UTD-MHAD
right arm swipe right	MAD	UTD-MHAD	draw x	UOW-Action 3D	UTD-MHAD
baseball swing from left	MAD	UTD-MHAD			

Table 1: Combined-15 action dataset.

15 actions selected from six public datasets including MSRC-12 [48], WorkoutSU-10 [49], Multi-modal action database (MAD) [51], UTD Multi-modal human action dataset (UTD-MHAD) [52], UT-Kinect [53] and UOW-Action 3D [54]. Samples of each action were from two of the six datasets and performed twice by 16 subjects (8 subjects in each dataset). In total, there are 480 action instances and 48 different subjects. Table 1 lists action names and their source datasets. All actions can be texturally decribed by using 19 semantic poses.

Combined-17 dataset: This dataset was specifically created to evaluate zero-shot action recognition by combining 17 actions selected from six datasets including MSRC-12 [48], WorkoutSU-10 [49], UT-Kinect [53], UOW-Action 3D [54], UCF-Kinect [55] and WorkoutUOW-18. Table 1 shows the action names and their source datasets. Two samples each of ten subjects for every action were randomly selected to construct the dataset. In total, there are 340 action instances and 60 different subjects. There are 20 semantic poses were parsed from the textual instructions of the 17 actions.

5.2. Extraction of salient frames

The main purpose of extracting salient frames rather than using all frames is to reduce the computation and the extraction acts like a non-uniform sampling. To this end, it is required that the salient frames should well represent the frames on the original sequence. MSRC-12 and Combined-17 datasets were used to evaluate the representativeness of salient frames extracted based on potential energy (PE), kinetic energy (KE) and total energy (TE). The use of Combined-17 is to avoid the possible bias on individual datasets.

Action	Source dataset	Action	Source dataset
sit to stand (1)	UT-Kinect	beat (10)	MSRC-12
stand to sit (2)	UT-Kinect	hip flexicon (11)	WorkouSU-10
push (3)	UT-Kinect	hip adductor stretch (12)	WorkoutSU-10
pull (4)	UT-Kinect	left lunge with right kick (13)	WorkoutUOW-18
draw x (5)	UOW-Action 3D	right lunge with knee up (14)	WorkoutUOW-18
draw tick (6)	UOW-Action 3D	standing clamshells (15)	WorkouUOW-18
right leg side kick (7)	UOW-Action 3D	standing triceps extension (16)	WorkouUOW-18
lift arms (8)	MSRC-12	side to side bend (17)	WorkouUOW-18
had enough (9)	MSRC-12		

Table 2: Combined-17 action dataset.

The representativeness of salient frames was measured using the sparse reconstruction errors of actions samples by treating the salient frames as the bases of a dictionary. Suppose an action instance is written as $\mathbf{Y}$ and salient frames are represented by $\mathbf{D}$ with each column being a salient frame (basis), the orthogonal matching pursuit [56] algorithm was used to obtain the coefficient $\mathbf{C}$ with a fixed sparsity and the reconstruction error was calculated through $||\mathbf{Y} - \mathbf{DC}||_F^2$. Figure 6 shows the average reconstruction error of all action instances in the two datasets with the sparsity of being 3. It can be seen that TE produced the smallest error while PE produced the largest error. In other words, salient frames extracted using TE are better representative than the frames extracted using PE and KE individually. It is observed the number of extracted salient frames is on average only about $6\% - 15\%$ of the number of frames in an action instance depending on the complexity of the action, which substantially reduces the computation later in the proposed method.

5.3. Visual pose modelling

Experiments were conducted on MSRC-12 and Combined-17 datasets to evaluate the discrimination of the visual poses. Considering visual poses naturally form a codebook, the classical bag-of-words method was adopted to generate action descriptors and an SVM was trained for classification. Specifically, half of subjects in the datasets were randomly selected and their samples were used for training. Salient frames were extracted from the training samples based on TE and a GMM was fitted to the moving pose descriptors of the salient frames, each Gaussian component being considered to

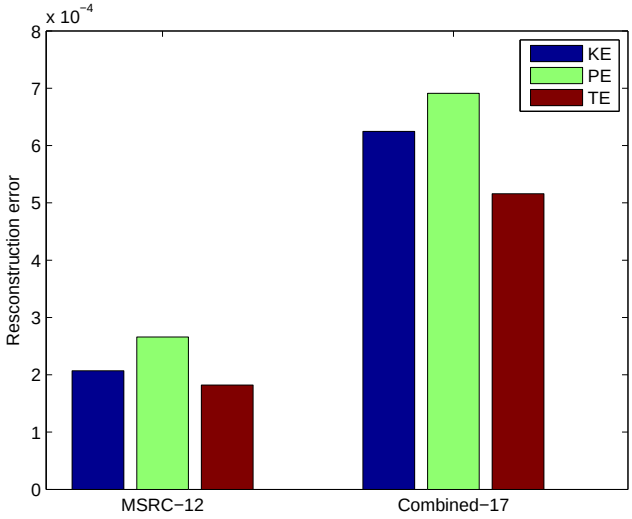
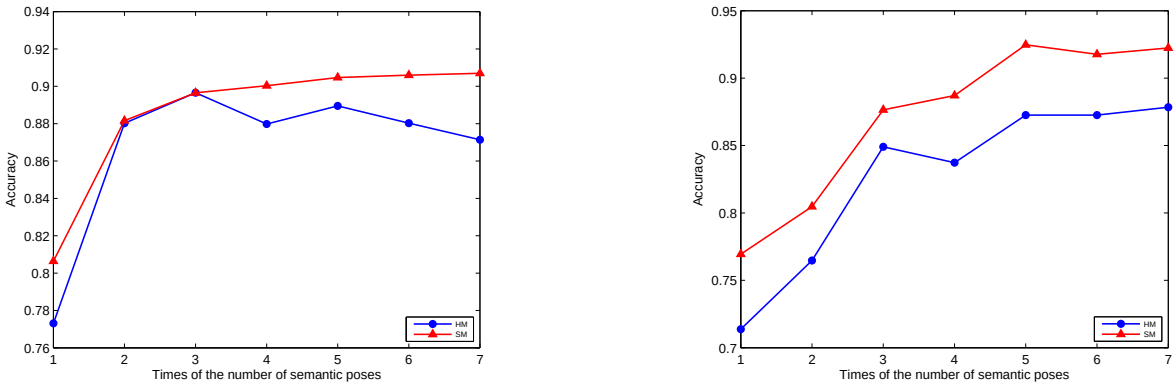


Figure 6: Average sparse reconstruction errors of action samples using as the dictionaries salient frames extracted based on PE, KE and TE respectively.

be a visual pose and treated as a visual word. A histogram of the visual words mapped from the salient frames of an action instance is constructed as the action descriptor for classification. Two mappings were evaluated, one is soft mapping (SM) which votes the histogram using the probability of a salient frame belonging to the visual poses or words, the other is hard mapping (HM) which votes the histogram by classifying a salient frame into one of the visual words exclusively. Figure 7a and 7b respectively show the classification accuracies versus with number of visual poses on the MSRC-12 and Combined-17 datasets where the number of visual poses was set to multiple times of the number semantic poses. Notice that the classification accuracy increases with the increasing number of visual poses and then becomes saturated after the number of visual poses is 3 or more times as many as that of semantic poses. It can also be noticed that SM has a clear advantage over HM regardless of the number of visual poses.

5.4. Lexicon learning

This section demonstrates how visual poses are probabilistically associated with semantic poses in a learned lexicon. Taking the Combined-17 dataset as an example, salient frames were extracted based on TE from (training) samples of randomly selected half of the subjects and visual poses were learned by fitting a GMM with different number of components. Figure 8 shows some sample semantic poses (shaded circles) of the learned lexicon, the top three most associated visual poses and their probabilities (values on the links), where "2" refers to the semantic pose of "Outstretched arms are above the shoulder", "12" is the semantic pose of "The right thigh is at an angle to the corresponding shin with



(a) MSRC-12 dataset

(b) Combined-17 dataset

Figure 7: Classification accuracies v.s number of visual poses.

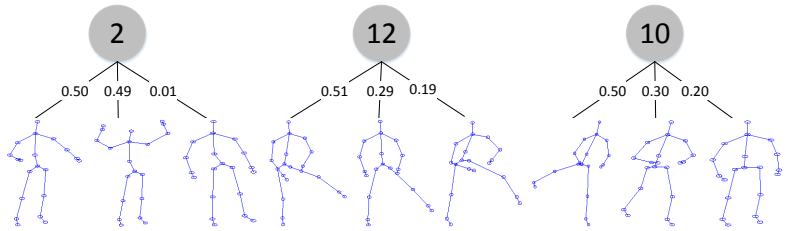
outstretched left leg" and "10" represents "Raised and outstretched right leg is on the right side of the body". From the learned lexicons with different number of visual poses, it is observed that

(1) Overall, the visual poses corresponds to the semantic pose, i.e. textural description, reasonably well. There are noticeable advantages of allowing the association of multiple visual poses with one semantic pose, including being insensitive to performing styles of same action by different subjects and variances of the same semantic pose in different actions or sequences of semantic poses, i.e. context. For example, semantic pose "2" is included in action "lift arms". Arms go through the torso as shown in the first visual pose linked to "2" of Figure 8a when reaching "2" from its previous semantic pose "arms are beside the body".

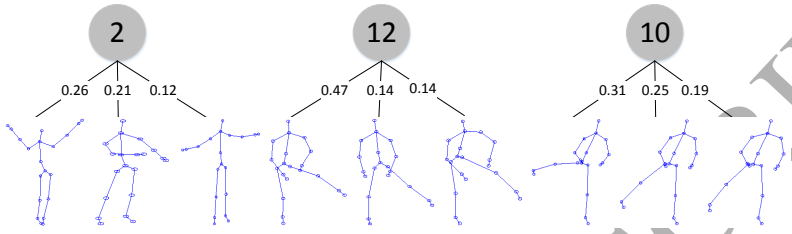
(2) As expected, more visual poses would be associated with each semantic pose as the number of visual poses increases. This is indicated by the accumulated probabilities of the top three visual poses in Figures 8a, 8b and 8c. The accumulated probabilities (larger than 0.9) in Figure 8a are higher than those in Figure 8b and 8c where the numbers of visual poses are larger. In general, more number of visual poses would improve the tolerance to visual variance within individual semantic poses. However, the tolerance would not be improved further when the number of visual poses reaches a certain value.

5.5. Classification and Comparisons

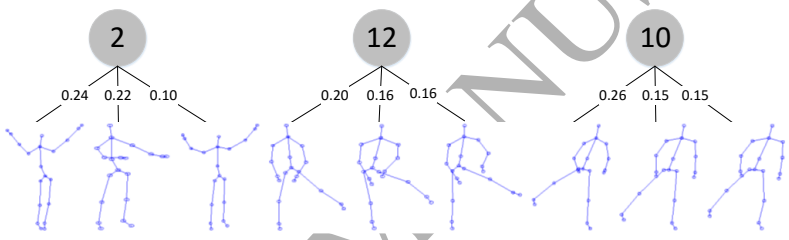
This section presents classification results of the proposed method on the five datasets. Three protocols, cross-subject, cross-dataset and zero-shot, were used in the evaluation.



(a) Number of visual poses is same as the number of semantic poses



(b) Number of visual poses is 3 times as many as the number of semantic poses



(c) Number of visual poses is 6 times as many as the number of semantic poses

Figure 8: Sample semantic poses, skeletons of their top three associated visual poses and the probabilities in learned lexicons with different number of visual poses.

5.5.1. Cross-subject test

In cross-subject test, the proposed method was evaluated on all of the five datasets. For each dataset, samples of randomly selected half of subjects were chosen for training and samples of the rest subjects were used for testing. Confusion matrices and learned visual poses of the five datasets are presented together with a comparison to a baseline and state-of-the-art methods.

Classification accuracy: For the MSRC-12 dataset, the proposed method achieved 92.03%. From the confusion matrix shown in Figure 9a, it can be seen that classification accuracy for individual actions are all over 80%. On the WorkoutSU-10 dataset, the proposed method achieved 99.00% accuracy and only 5 instances from actions "hip flexion", "trunk rotation" and "lateral stepping" were misclassified as shown in Figure 9b. The accuracy of the WorkoutUOW-18

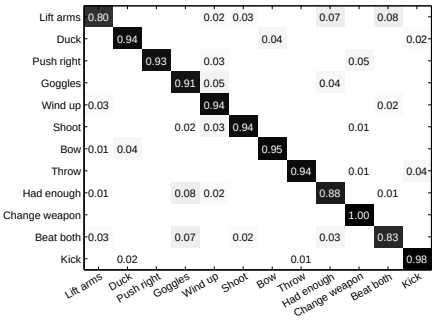
dataset is 94.97% and its confusion matrix is shown in Figure 9e. Most of the actions can be recognized with 100% accuracy, but some actions appear difficult to discriminate, such as "bent over row" and "bent over lateral raise". This is probably because the two actions have same torso movement and confused arm movements. On the Combined-15 dataset, the classification accuracy is 93.33%. The confusion matrix shown in Figure 9c indicates that action "cross arms" was misclassified into "clap hands", which is reasonable because their visually poses are hardly discriminative (e.g. arms are close to chest). The proposed method achieved 94.12% accuracy on the Combined-17 dataset. As indicated in the confusion matrix of this dataset shown in 9d, action "beat" is confused with action "lift arms" and "had enough" due to similar visual poses.

Figure 10 shows classification accuracies of the five datasets versus the number of visual poses where TE was used to extract salient frames. The number of visual poses was set to one to seven times as many as the number of semantic poses. As seen in Figure 10, as the number of visual poses increases, classification is first improved significantly and then levelled at a plateau. This is consistent with Figure 7. When the number of visual poses is too small, the learned lexicon would be sensitive to intra-class variations due to action types and context of the semantic poses as discussed above.

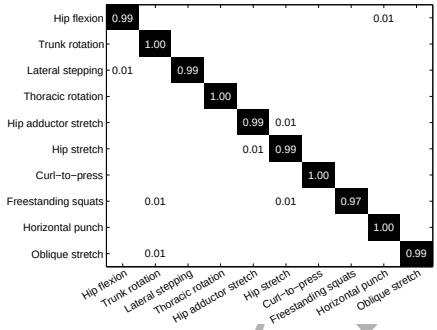
Comparison with the baseline: The baseline methods are built upon the traditional hidden Markov alignment model (HMAM) [34]. Specially, the proposed model is compared with HMAM combining with two variations: hard mapping (HM), i.e. mapping salient frames to their most likely visual poses exclusively, hard mapping with temporal constraint (HM + T) on the alignment of a visual pose sequence with a semantic pose sequence. In both variations, visual pose sequences are first generated through the hard mapping and pose lexicon is then learned by aligning visual pose sequences with semantic pose sequences as detailed in [20]. Table 3 summarizes the cross-subject test results of the HM, HM+T and proposed methods.

The baseline HM+T outperforms the baseline HM in almost all cases significantly, specifically by about 12 percentage points on the Combined-15 dataset and 20 percentage points on the Combined-17 dataset. This is mainly because that HM fails to distinguish different temporal orders in the sequences of semantic poses, such as "stand to sit" and "sit to stand", "push" and "pull".

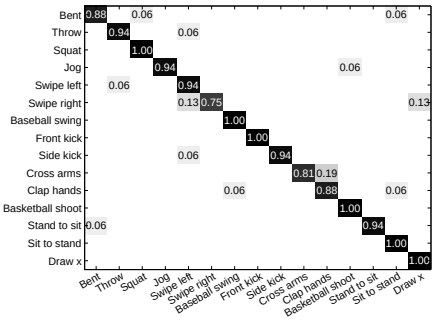
The proposed model further improves the classification accuracies comparing with the baseline HM+T. This improvement demonstrates the advantages of joint learning of hidden visual pose sequences and alignment sequences. In particular, it can improve the classification of actions with large intra-variance often introduced by the movement of many body parts. For example, the action "wind up" in the MSRC-12 dataset involves different body parts such as hands,



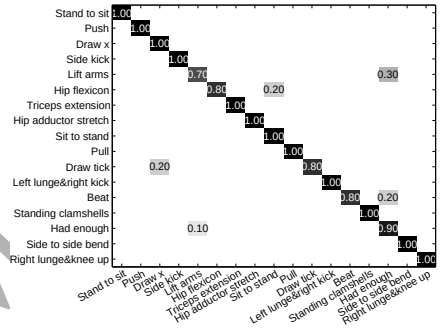
(a) MSRC-12 dataset.



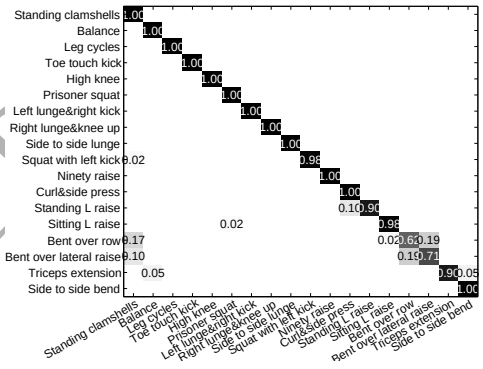
(b) WorkoutSU-10 dataset.



(c) Combined-15 dataset.



(d) Combined-17 dataset.



(e) WorkoutUOW-18 dataset.

Figure 9: Confusion matrices of cross-subject tests.

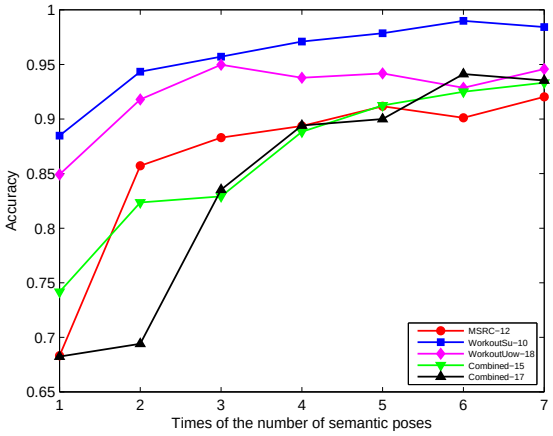
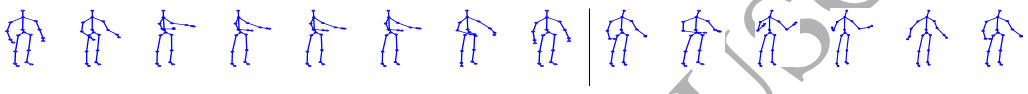


Figure 10: Recognition accuracy versus the number of visual poses.

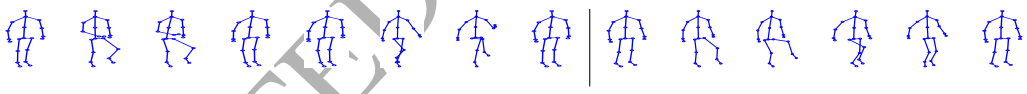


(a) Wind up with hands

(b) Wind up with forearms

(c) Wind up with whole arms

Figure 11: Example instances of action “wind up”.



(a) High knee with arms swing

(b)High knee without arms swing

Figure 12: Example instances of action “high knee”.

forearms or whole arms (see Figure 11). Its recognition rate is improved by about 10 percentage points by the proposed model using TE. Similar situation exists for the action “high knee” in the WorkoutUOW-18 dataset which can be performed with or without arms swing (see Figure 12). Furthermore, recognition of actions with small inter-variance has also been improved. For example, the recognition of easy-to-be-confused actions “cross arms” and “clap hands” have been improved by 12.5 and 6.25 percentage points respectively by the the proposed method with TE. The results also confirm the advantage of TE over PE and KE in extracting salient frames.

Methods		HM	HM + T	Proposed Method
MSRC-12	KE	90.08%	90.9%	**90.91%**
	PE	86.99%	86.39%	**89.58%**
	TE	90.34%	91.75%	**92.03%**
WorkoutSU-10	KE	**98.29%**	97.43%	97.57%
	PE	96.43%	97%	**98.00%**
	TE	97.88%	98.71%	**99.00%**
WorkoutUOW-18	KE	88.76%	91.14%	**93.78%**
	PE	91.01%	90.35%	**92.33%**
	TE	93.25%	93.78%	**94.97%**
Combined-15	KE	80%	92.5%	**92.92%**
	PE	75.42%	86.67%	**88.75%**
	TE	81.25%	92.08%	**93.33%**
Combined-17	KE	68.82%	94.12%	**94.12%**
	PE	69.41%	93.53%	**93.53%**
	TE	72.94%	93.53%	**94.12%**

Table 3: Classification results of the proposed method and comparison with the two variations of the baseline. Note that PE, KE and TE are respectively the criteria used to extract salient frames.

Comparison with other methods: The proposed method is also compared with two categories of state-of-the-art methods. The first category is the methods using skeleton data that achieved the state-of-the-art results on MSRC-12 and WokroutSU-10 datasets, including covariance descriptor method (Cov3DJ) [50], random decision forest (RDF) [49] and extended LC-KSVD [21]. To demonstrate the performance of semantic action recognition of the proposed method, the second category is the state-of-the-art semantic learning methods such as the classical latent Dirchlet allocation (LDA) [12]. Visual poses and semantic poses in the proposed method were considered respectively as the words and topics in LDA. After action was represented by a vector of topics, a linear-SVM was trained for classification. To further demonstrate the advantages of using semantic pose sequences, the proposed method is compared with the traditional hidden Markov model (HMM) [57]. One HMM was trained for each action independently by taking video frame sequences

Methods	RDF [49]	Cov3DJ [50]	LDA [12]	HMM [57]	roposed Method
MSRC-12	**94.03%**	91.70%	90.25%	88.50%	92.03%
WorkoutSU-10	98.00%	-	98.43%	95.86%	**99.00%**
WorkoutUOW-18	-	84.79%	93.12%	76.46%	**94.97%**
Combined-15	-	74.17%	82.92%	82.08%	**93.33%**
Combined-17	-	70.00%	84.12%	80.00%	**94.12%**

Table 4: Cross-subject classification accuracies of the proposed method and other state-of-the-art methods.

only as the input and setting the number of hidden states same as the number of semantic poses. Table 4 reports the results. As seen, the proposed method outperforms all other methods on WorkoutSU-10, WorkoutUOW-18, Combined-15 and Combined-17 datasets. In the case of MSRC-12 dataset, the proposed method achieved the second best result, with 2 percentage points lower than that of RDF. Notice that the proposed method achieved a better result than LDA and offers an effective semantic action recognition. Moreover, the proposed method outperformed the traditional HMM method by between 3 to 19 percentage points on these datasets.

5.5.2. *Cross-dataset test*

Experiments were conducted using a cross-dataset protocol to evaluate the proposed method on the Combined-15 dataset. In particular, samples in the dataset was divided into two parts according to their source: samples from source dataset 1 and 2 as shown in Table 1, one part was used as training and the other part was used as test. Two tests were conducted by alternating the training and test samples. Table 5 shows the classification results. We compared the proposed method with a state-of-the-art transfer learning method, called joint distribution adaptation (JDA) [58], because it has performed well in cross-dataset object recognition. From Table 5 it can be seen that the proposed method achieved comparable results to JDA although it is not specially designed for transfer learning. This indicates that the proposed method is not as sensitive to the environmental settings of different datasets as most action recognition methods [59, 60] would do, one of the key advantages brought by semantic action recognition.

5.5.3. *Zero-shot test*

The Combined-17 dataset was used to evaluate the proposed method under a zero-shot test protocol where no instances of testing actions were included in training. Based on the actions and their semantic poses in the Combined-17 dataset,

Training	Test	Methods	Accuracy
Source dataset 1	Source dataset 2	JDA [58]	76.25%
		Proposed	76.67%
Source dataset 2	Source dataset 1	JDA [58]	74.17%
		Proposed	74.21%

Table 5: Results of cross-dataset classification.

five valid pairs of training-testing action sets were generated. Being valid, it means that the semantic poses of the testing actions are a subset of the semantic poses of the training actions. For each valid pair of training-testing action sets, all instances of the training actions, i.e. 20 instances per action, were used to train the pose and lexicon models and classification was performed on all the instances of testing actions. Table 6 shows the five valid pairs of training-testing action sets and their classification accuracies of individual actions and average accuracies over the testing actions. In the table, actions are numbered and their names are shown in Table 2. It is seen that more than half of actions in the five cases were well recognized with an accuracy over 90%, which demonstrates that the proposed method is effective in recognizing novel actions. This experiment has demonstrated the advantages of semantic pose sequences in action recognition and semantic based action recognition is a promising approach to zero-shot recognition.

5.6. Discussion

The current implementation of the algorithm assumes that actions are performed completely. However, forgotten parts of an action may happen at the beginning or ending, but less likely in the middle, of the action as an action is usually a continuous movement of human body. Extension of the algorithm to deal with the forgotten parts is possible and not difficult. One option is to expand X to "NULL+salient frames+NULL", where "NULL" represents the forgotten parts with a constraint on how many parts can be missed.

6. Conclusion and future works

This paper presents a novel method for semantic action recognition through a pose lexicon and extends the hidden Markov alignment model to learn the lexicon. Given a sequence of semantic poses extracted from textual instructions of an action and a sequence of visual frames, the proposed model find the most likely hidden visual pose sequence from

Training actions	Recognition accuracies of the testing actions (%)								
1, 4, 6, 9, 10,	2	3	5	7	8	11	12	16	Average
13, 14, 15, 17	100.00	65.00	100.00	90.00	90.00	70.00	80.00	100.00	86.87
2, 3, 6, 9, 10,	1	4	5	7	8	11	12	16	Average
13, 14, 15, 17	90.00	100.00	100.00	95.00	75.00	100.00	100.00	90.00	93.75
2, 4, 5, 9, 10,	1	3	6	7	8	11	12	16	Average
13, 14, 15, 17	100.00	85.00	100.00	100.00	85.00	100.00	100.00	70.00	92.5
1, 3, 5, 9, 10,	2	4	6	7	8	11	12	16	Average
13, 14, 15, 17	100.00	85.00	100.00	95.00	85.00	100.00	100.00	65.00	91.25
2, 3, 5, 9, 10,	1	4	6	7	8	11	12	16	Average
13, 14, 15, 17	100.00	85.00	100.00	95.00	100.00	100.00	100.00	90.00	96.25

Table 6: Results of zero-shot recognition. In the table, actions are numbered and their corresponding names are shown in Table 2.

low-level visual features and learns an alignment with the semantic pose sequence by enforcing a temporal constraint that is essential in classifying actions. Experimental evaluation has been conducted on skeleton data on five datasets and the results have verified the efficacy of the proposed method. The proposed method can be extended in a number of ways to improve its robustness, including use of other modalities such as RGB and/or depth to model the visual poses, adoption of deep learning scheme to learn the visual features and visual pose models, body parts-based semantic action recognition and improvement of the semantic poses to distinguish subjects(figure) and objects(ground) for object interactions.

In addition, the learned lexicon can also be used for generating semantic description of actions. Such a problem can be formulated as finding the most likely semantic pose sequence given a video frame sequence X. One approach is to perform the task in two steps: (1) generate all semantic pose sequence candidates given the semantic pose set Ψ (where the set of semantic pose sequences is written as Φ), (2) find the optimal semantic pose sequence ϕ that maximizes $P(X|\phi)$ from Φ. Step (2) can be solved by the algorithm presented in Section 3.2. Step (1) is basically a search problem and algorithms (i.e. greedy decoding [61]) need to be developed to reduce the search space.

References

[1] W. Li, Z. Zhang, Z. Liu, Expandable data-driven graphical modeling of human actions based on salient postures, IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1499–1510.

[2] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Comput. Vis. Image Underst. 115 (2) (2011) 224–241.

[3] P. Wang, W. Li, P. Ogunbona, Z. Gao, H. Zhang, Mining mid-level features for action recognition based on effective skeleton representation, in: Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2014, pp. 1–8.

[4] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, P. Ogunbona, Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring, in: Proceedings of the 23rd ACM international conference on Multimedia (ACMMM), 2015, pp. 1119–1122.

[5] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, P. O. Ogunbona, Action recognition from depth maps using deep convolutional neural networks, IEEE Trans. Hum. Mach. Syst. 46 (4) (2016) 498–509.

[6] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), 2010, pp. 9–14.

[7] C. Wang, Y. Wang, A. L. Yuille, An approach to pose-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 915–922.

[8] A. Eweiwi, M. S. Cheema, C. Bauckhage, J. Gall, Efficient pose-based action recognition, in: Proceedings of the 12th Asian Conference on Computer Vision (ACCV), Springer, 2014, pp. 428–443.

[9] L. Talmy, Lexicalization patterns: Semantic structure in lexical forms, Lang. Typology and Syntactic Description 3 (1985) 57–149.

[10] S. Petrov, L. Barrett, R. Thibaux, D. Klein, Learning accurate, compact, and interpretable tree annotation, in: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics(Coling ACL), 2006, pp. 433–440.

[11] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[12] J. C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, Int. J. Comput. Vision 79 (3) (2008) 299–318.

[13] C. Wu, J. Zhang, S. Savarese, A. Saxena, Watch-n-patch: Unsupervised understanding of actions and relations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4362–4370.

[14] J. Liu, B. Kuipers, S. Savarese, Recognizing human actions by attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3337–3344.

[15] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, B. Schiele, Script data for attribute-based recognition of composite activities, in: Proceedings of the 12th European Conference on Computer Vision (ECCV), 2012, pp. 144–157.

[16] Z. Moghaddam, M. Piccardi, Training initialization of hidden markov models in human action recognition, IEEE Trans. Automation Science Engineering 11 (2) (2014) 394–408.

[17] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, K. Saenko, Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2712–2719.

[18] S. Sadanand, J. J. Corso, Action bank: A high-level representation of activity in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1234–1241.

[19] I. Lillo, A. Soto, J. Carlos Niebles, Discriminative hierarchical modeling of spatio-temporally composable human activities, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 812–819.

[20] L. Zhou, W. Li, P. Ogunbona, Learning a pose lexicon for semantic action recognition, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2016, pp. 1–6.

[21] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, H. Zhang, Discriminative key pose extraction using extended

LC-KSVD for action recognition, in: Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2014, pp. 1–8.

[22] Y. Wang, G. Mori, Human action recognition by semi-latent topic models, IEEE Trans. Pattern Anal. Mach. Intell. 31 (10) (2009) 1762–1774.

[23] S. Yang, C. Yuan, W. Hu, X. Ding, A hierarchical model based on latent dirichlet allocation for action recognition, in: Proceedings of the 22nd International Conference on Pattern Recognition (ICPR), 2014, pp. 2613–2618.

[24] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (1-2) (2001) 177–196.

[25] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[26] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, Attribute regularization based human action recognition, IEEE T. Inf. Foren. Sec. 8 (10) (2013) 1600–1609.

[27] C. Liu, X. Wu, Y. Jia, A hierarchical video description for complex activity understanding, Int. J. Comput. Vision (2016) 1–16.

[28] T. S. Motwani, R. J. Mooney, Improving video activity recognition using object recognition and text mining, in: Proceedings of the 20th European Conference on Artifical Intelligence (ECAI), 2012, pp. 600–605.

[29] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, S. Guadarrama, Generating natural-language video descriptions using text-mined knowledge, in: Proceedings of the 27th AAAI Conference on Aritificial Intelligence (AAAI), 2013, pp. 541–547.

[30] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, B. Schiele, Translating video content to natural language descriptions, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 433–440.

[31] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, K. Saenko, Translating videos to natural language using deep recurrent neural networks, The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).

[32] R. Xu, C. Xiong, W. Chen, J. J. Corso, Jointly modeling deep video and compositional text to bridge vision and language in a unified framework, in: Proceedings of the 29th AAAI conference on Aritificial Intelligence (AAAI), 2015, pp. 2346–2352.

[33] I. Naim, Y. C. Song, Q. Liu, H. Kautz, J. Luo, D. Gildea, Unsupervised alignment of natural language instructions with video segments, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI), 2014, pp. 1–8.

[34] S. Vogel, H. Ney, C. Tillmann, HMM-based word alignment in statistical translation, in: Proceedings of the 16th Conference on Computational Linguistics (Coling), 1996, pp. 836–841.

[35] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, R. L. Mercer, The mathematics of statistical machine translation: Parameter estimation, Comput. Ling. 19 (2) (1993) 263–311.

[36] I. Naim, Y. C. Song, Q. Liu, L. Huang, H. Kautz, J. Luo, D. Gildea, Discriminative unsupervised alignment of natural language instructions with corresponding video segments, in: Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT), 2015, pp. 164–174.

[37] C. Dyer, J. Clark, A. Lavie, N. A. Smith, Unsupervised word alignment with arbitrary features, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT), 2011, pp. 409–419.

[38] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, C. Schmid, Weakly-supervised alignment of video with text, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4462–4470.

[39] Y. C. Song, I. Naim, A. Al Mamun, K. Kulkarni, P. Singla, J. Luo, D. Gildea, H. Kautz, Unsupervised alignment of actions in video with text descriptions, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 1–7.

[40] E. Charniak, Immediate-head parsing for language models, in: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL), 2001, pp. 124–131.

[41] M. Zhu, Y. Zhang, W. Chen, M. Zhang, J. Zhu, Fast and accurate shift-reduce constituent parsing, in: Proceedings of the annual meeting of the Association for Computational Linguistics (ACL), 2013, pp. 434–443.

[42] D. Chen, C. D. Manning, A fast and accurate dependency parser using neural networks, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 740–750.

[43] R. McDonald, K. Crammer, F. Pereira, Online large-margin training of dependency parsers, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL), 2005, pp. 91–98.

[44] B. Bohnet, Very high accuracy and fast dependency parsing is not a contradiction, in: Proceedings of the 23rd International Conference on Computational Linguistics (Coling), 2010, pp. 89–97.

[45] J. Nivre, An efficient algorithm for projective dependency parsing, in: Proceedings of the 8th International Workshop on Parsing Technologies (IWPT), 2003, pp. 149–160.

[46] F. J. Och, H. Ney, A systematic comparison of various statistical alignment models, Computat. Ling. 29 (1) (2003) 19–51.

[47] M. Zanfir, M. Leordeanu, C. Sminchisescu, The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2752–2759.

[48] S. Fothergill, H. Mentis, P. Kohli, S. Nowozin, Instructing people for training gestural interactive systems, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (SIG CHI), 2012, pp. 1737–1746.

[49] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, A. Erçil, A decision forest based feature selection framework for action recognition from RGB-Depth cameras, in: Image Anal. Recognit., Springer, 2013, pp. 648–657.

[50] M. E. Hussein, M. Torki, M. A. Gowayyed, M. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptor on 3D joint locations, in: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), 2013, pp. 2466–2472.

[51] D. Huang, S. Yao, Y. Wang, F. De La Torre, Sequential max-margin event detectors, in: Proceedings of the 13th European Conference on Computer Vision (ECCV), 2014, pp. 410–424.

[52] C. Chen, R. Jafari, N. Kehtarnavaz, Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2015, pp. 168–172.

[53] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 20–27.

[54] C. Tang, W. Li, C. Hou, P. Wang, Y. Hou, J. Zhang, P. Ogunbona, Online action recognition based on incremental learning of weighted covariance descriptors, arXiv preprint arXiv:1511.03028.

[55] H. Boyraz, S. Masood, B. Liu, M. Tappen, H. Foroosh, Action recognition by weakly-supervised discriminative region localization, in: Proceedings of the British Machine Vision Conference (BMVC), 2014, pp. 1–13.

[56] Y. C. Pati, R. Rezaiifar, P. Krishnaprasad, Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition, in: Proceedings of the Conference Record of The Twenty-Seventh Asilomar Conference on the Signals, Systems and Computers, 1993, pp. 40–44.

[57] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[58] M. Long, J. Wang, G. Ding, J. Sun, P. S. Yu, Transfer feature learning with joint distribution adaptation, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2200–2207.

[59] J. Zhang, W. Li, P. Wang, P. Ogunbona, S. Liu, C. Tang, A large scale rgb-d dataset for action recognition, in: Proceedings of International Conference on Pattern Recognition Workshop on UHA3DS (ICPRW), 2016.

[60] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, C. Tang, Rgb-d-based action recognition datasets: A survey, Pattern Recogn. 60 (2016) 86–105.

[61] U. Germann, Greedy decoding for statistical machine translation in almost linear time, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL HLT), 2003, pp. 1–8.

Lijuan Zhou: received her bachelor and master degrees in software engineering from Zhengzhou University, in 2009 and 2012, respectively. She is currently a PhD candidate student in the Advanced Multimedia Research Lab of University of Wollongong, Australia. Her research interests include human action recognition and action understanding.

Wanqing Li: received his PhD in electronic engineering from University of Western Australia. He is an Associate Professor and Co-Director of the Advanced Multimedia Research Lab of University of Wollongong, Australia. His research areas are machine learning, 3D computer vision, 3D multimedia signal processing and medical image analysis.

Philip Ogunbona: earned his PhD, DIC (Electrical Engineering) from Imperial College, London. His Bachelor degree was in Electronic and Electrical Engineering from University of Ife, Nigeria. He is a Professor and Co-Director of Advanced Multimedia Research Lab, University of Wollongong, Australia. His research interests include computer vision, pattern recognition and machine learning.

Zhengyou Zhang: received his PhD and D.Sc. in computer science from University of Paris XI. He is a Principal Researcher, Research Manager of the MIX Group with Microsoft Corporation, IEEE and ACM Fellow. His research interests include computer vision, speech processing, multisensory fusion, multimedia computing, and humanmachine interaction.