Discriminative Hierarchical Modeling of Spatio-Temporally Composable Human Activities

Ivan Lillo, Alvaro Soto
P. Universidad Catolica de Chile.
ialillo@uc.cl, asoto@ing.puc.cl

Juan Carlos Niebles
Universidad del Norte. Colombia.
njuan@uninorte.edu.co

Abstract

This paper proposes a framework for recognizing complex human activities in videos. Our method describes human activities in a hierarchical discriminative model that operates at three semantic levels. At the lower level, body poses are encoded in a representative but discriminative pose dictionary. At the intermediate level, encoded poses span a space where simple human actions are composed. At the highest level, our model captures temporal and spatial compositions of actions into complex human activities. Our human activity classifier simultaneously models which body parts are relevant to the action of interest as well as their appearance and composition using a discriminative approach. By formulating model learning in a max-margin framework, our approach achieves powerful multi-class discrimination while providing useful annotations at the intermediate semantic level. We show how our hierarchical compositional model provides natural handling of occlusions. To evaluate the effectiveness of our proposed framework, we introduce a new dataset of composed human activities. We provide empirical evidence that our method achieves state-of-the-art activity classification performance on several benchmark datasets.

1. Introduction

Humans are capable of performing multiple simple tasks simultaneously. In particular, humans can control the motion of their limbs and torso with tremendous precision, and can associate subsets of their limbs to different tasks. People can, for example, text while walking, or wave one hand while holding a phone to their ear with the other (Fig. 1). It is interesting to note that different compositional arrangements of simple body motions can yield different semantics at a higher level. Compositions can occur spatially and/or temporally; for example, in a particular sport, the referee can signal different events by raising the left hand instead of the right hand (spatial composition) or by executing a specific sequence of gestures (temporal composition).

Figure 1. People perform complex activities that can be characterized as spatial and/or temporal compositions of simpler actions. **Top-left:** A person simultaneously *waves* and *walks* by assigning subsets of body parts to different actions. **Top-right:** A person sequencially *talks on the phone* and *runs* away to attend an urgent matter. **Bottom:** A person *walks* in a room, *picks a book up*, *walks* while *reading a book*, etc. In this paper, we propose a novel formulation that is able to capture these spatio-temporal compositions for complex activity recognition using RBGD data.

This paper proposes a computational framework for modeling complex activities by capturing their spatio-temporal composition. We achieve this by introducing a unified hierarchical model that operates at three semantic levels: at the bottom level, our model builds a discriminative dictionary of body pose primitives; at the mid-level, these poses are combined to compose atomic actions; finally, at the top-level, atomic actions are combined to compose complex activities.

In our model, we formulate learning as an energy minimization problem, where structural hierarchical relations are modeled by sub-energy terms that model and connect the different abstraction levels: poses, actions, and activities. By using a multi-class max-margin approach and coupling learning of all abstraction levels, we are able to obtain discriminative and functional mid-level representations that foster pose sharing among action classes and action shar-

ing among target activities. This allows us to use small diccionary sizes, to reduce overfitting problems, and to improve the computational efficiency.

From a machine learning perspective, the use of mid-level representations based on body poses and atomic actions with clear semantic meaning, facilitates the acquisition of labeled trained data. In particular, at training time high level semantic information at the level of activities and atomic actions is propagated down the hierarchy to guide the otherwise unsupervised search for relevant primitives at the level of body poses. We believe this is a powerful learning framework that provides a rich hypothesis space to build visual compositional schemes [4]. This also marks a notable difference with respect to current popular compositional hierarchical models based on deep learning techniques [3], where the lack of easy interpretation of the resulting mid-level representations complicates the training process.

In fact, the spatio-temporal compositional abilities of our model also bring other conceptual advantages. First, our framework is capable of identifying the active body parts when different actions are executed, while also capturing their spatial appearance and temporal evolution. Second, our model can naturally handle scenarios of partial occlusions and pose estimation failures by inferring an appropiate spatial composition that assigns higher weight to the visible/relevant body parts while dismissing the occluded/irrelevant ones. We show empirical evidence of these properties in our experiments.

The rest of the paper is organized as follows. First, we review some of the relevant prior work in Sec. 2. In Sec. 3 we introduce the details of our model and discuss the learning and inference algorithms. Sec. 4 presents empirical evaluations of our method in a new benchmark dataset and Sec. 5 presents our main conclusions.

2. Related Work

There is a large body of literature related to human activity recognition as, it is one of the most active topics in computer vision. We refer the reader to [1] for an extensive review of recent advances in the field. Here, we discuss some of the most relevant prior work related to our paper.

A number of researchers have tackled the problem of temporal composition of actions using representations based on local interest points [12, 19] by modeling their temporal arrangement. Some researchers have extended image representations, such as correlatons [28] and spatial pyramids [21] to videos [20], and have applied them to the problem of simple human action categorization. Other researchers have proposed models for decomposing actions into short temporal motion segments [14, 25], but cannot capture spatial composition of actions. Recently, several graph-based models have been proposed to account for spatio-temporal composition of low-level features [2, 7, 8].

Instead of focusing on low level video features, our model builds on a pose-based representation, by first extracting information about the pose of the actor. There is a significant amount of pose-based action recognition methods in the literature. However, traditional methods have several limitations: sillouhette based recognition assumes the camera is static [5, 31]; methods based on 2D body configurations require expensive body part detections to obtain initial input; etc. Usually, even if accurate body pose estimation is available, these methods are tremendously affected by body part occlusions and by unrelated limb postures and motions that are not involved in the action. In [15], the authors propose a search engine for composed actions based on HMMs, but its application to activity classification is not discussed. Other line of work looks at annotating novel action videos by recognizing single actions [27], but ignoring the composition of those single actions into meaningful complex activities.

Related to our work, some propose pose-based action recognition models that leverage action compositions. In [17], a Markov Random Field is trained over small temporal segments, and includes object affordance labels and object detectors. In [36], wavelet features are computed over temporal segments in each body joint, and the model infers the underlying temporal structure of sequences of actions.

In this work, we avoid the difficulty and high computational expense of human pose estimation from color images and we rely on human body poses extracted from color+depth sequences. In particular, we employ the pose estimation algorithm from [29, 24]. Other researchers have also addressed the problem of activity recognition on color+depth data, but they usually focus on categorizing non-composed activities [30, 34, 38].

From a learning perspective, our work is related to methods for learning visual dictionaries from data. Early approaches were based on vector quantization, using k-means to cluster low-level keypoint descriptors [11]. These approaches spawned algorithmic variations that use alternative quantization methods, discriminative dictionaries, or pooling schemes [16, 21]. Recently, sparse coding methods have emerged as a powerful alternative to vector quantization providing dictionaries that achieve low reconstruction errors and attractive computational properties. For example, [9] achieves state-of-the-art performance when tested on several human action datasets. Discriminative sparse representations have also been proposed [6, 23], mostly building specific dictionaries for each class. In contrast to our approach, the literature focuses on non-hierarchical cases, where the dictionary construction considers only a specific abstraction level usually using a generative approach. Even in the case of learning a discriminative dictionary, there is usually only a weak conection between the dictionary construction and the implementation of top level classifiers.

Our model also builds on ideas related to learning classifiers using a discriminative framework and latent variables. In particular, [13] uses a latent SVM scheme to develop an object recognition approach based on mixtures of multiscale deformable part models. This model is later extended to the case of action recognition [25]. In contrast to our approach, the model in [13] is limited to binary classification problems. Recently, [35] proposes a hierarchical latent variable approach to action recognition that directly considers the multiclass classification case. The layered model in [35] incorporates information about patches, hidden-parts, and action class, where the meaning of the hidden layers is not clear. In contrast, our hierarchical model integrates semantically meaningful information at all layers: poses, actions and activities. Unlike [35], our model can account for compositions of actions into activities and, as a byproduct, outputs per-body-part and per-frame action classification, so it has the appealing property that mid-level semantics are produced in addition to the final activity classification decision. [34] proposes a model for action recognition in static images, so it solves a different problem. It is not clear if an extension to spatio-temporal compositions is possible. Similarly to our approach, [35] and [34] also use a latent svm machinery for model learning and inference, but details of the formulations are distinct to our framework. Furthermore, the novelty of our work is the proposed model, not the actual learning/inference algorithms.

In terms of hierarchical compositional models, our work is related to recent recognition approaches based on deep learning (DL) [3, 18], where the training process usually incorporates hierarchical estimation of latent variables, spatial pooling schemes, and intermediate representations based on linear filters. DL is usually applied over raw image representation using several layers of generic structures. As a consequence, DL architectures have many parameters and they are usually difficult to train. In contrast, we embed semantic knowledge to our model by explicitly exploiting compositional relations among poses, actions, and activities. This leads to simpler architectures and allows us to incorporate labeled data at intermediate layers. Furthermore, our max-margin approach is based on a Hinge loss, and not quadratic or logistic functions commonly used to train DL architectures leading to different optimization problems.

Our method tackles some limitations of previous work with a new framework that models spatio-temporal compositions of activities using a hierarchy of three sematic levels. The compositional properties of our model enable it to provide meaningful annotations and to handle occlusions naturally.

3. Model Description

We now introduce the details of our model. Fig. 2 summarizes our model graphically.

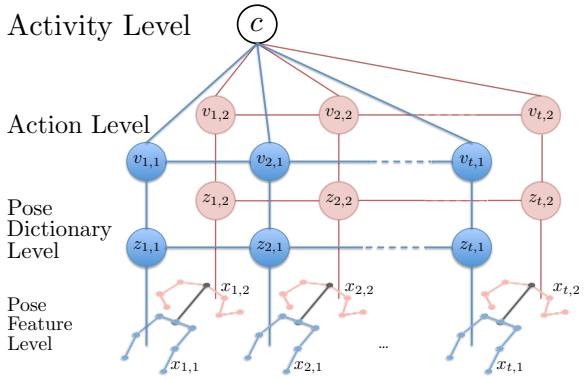


Figure 2. Overview of our discriminative hierarchical model for recognition of composable human activities. At the top level, activities are compositions of actions that are inferred at the intermedite level. These actions are in turn compositions of poses at the lower level, where pose dictionaries are learnt from data. Our model also divides each pose into R spatial regions to capture regions that are relevant to each activity. This figure illustrates the case when $R = 2$. Best viewed in color.

3.1. Video Representation

We represent actions as a sequence of human body poses estimated at each frame. Our algorithm extracts from RGB-D videos, a feature vector representing body poses using the methods in [29, 10]. Specifically, given a video D with T frames, we extract a feature vector $X = \{x_1, \ldots, x_T\}$, where x_t is a set of pose features extracted from the 3D body configuration estimated at frame t. Our pose features inspired by [10] include relative location between body joints, angles between limbs and angles between limbs and planes spanned by body parts.

3.2. Hierarchical Model

To recognize human activities and actions we propose a 3-level compositional hierarchical model. At the top level, our model assumes that each human activity is composed by a temporal and spatial arrangement of atomic actions. At the intermediate level, our model assumes that each atomic action is composed by a temporal arrangement of body poses. Finally, at the bottom level of the hierarchy, our model assumes that each body pose is composed by a spatial arrangement of features derived from RGB-D data. Given a video D, composed of T frames, where each frame t is described by a feature vector x_t, we define a video classification score, or energy function, for D as:

$$E(D) = E_{\text{activity}} + E_{\text{actions}} + E_{\text{poses}} + E_{\text{action transition}} + E_{\text{pose transition}}. \quad (1)$$

In Equation (1), energy E is expressed in terms of potentials associated to the activity present in video D and its related actions and poses. We also consider two energy potentials that encode information related to temporal transitions between pairs of consecutive actions and body poses. Next, we specify models for these energy terms.

At the lowest level of the hierarchy, the goal is to learn a dictionary of body poses using feature vectors x_t, $t \in [1,\dots,T]$. To achieve this, we introduce a latent vector $Z = (z_1 \dots z_T)$, where component z_t indicates the entry assigned to frame t from the dictionary of body poses. Let $w = (w_1 \dots w_K)$ be K coefficient vectors corresponding to a set of linear classifiers that define the entries of a dictionary of K visual poses. We define the energy term E_{poses} in Equation (1) as:

$$E_{\text{poses}} = \sum_{t=1}^{T} w_{z_t}^{\top} x_t = \sum_{t=1}^{T} \sum_{k=1}^{K} w_k^{\top} x_t \delta(z_t = k) \quad (2)$$

where $\delta(\ell) = 1$ if ℓ is true and $\delta(\ell) = 0$ if ℓ is false.

At the second level of the hierarchy, we use the dictionary of body poses to describe atomic actions using a bag-of-words (BoW) representation (average pooling). Specifically, $h^a(Z,V)$ is the histogram over the pose dictionary at those frames assigned to action a, where $V^{\top} = (v_1 \dots v_t \dots v_T)$ is a vector of action labels for each frame. Also, let $\beta_a = (\beta_{a,1}, \dots \beta_{a,K})$ be the K coefficients of a linear classifier associated to action a, $a \in [1,\dots,A]$. Each entry k in $h^a(Z,V)$ is given by:

$$h_k^a(Z,V) = \sum_{t=1}^{T} \delta(z_t = k)\delta(v_t = a) \quad (3)$$

Then, the energy potential of the action labels V for all frames of video D is given by:

$$E_{\text{action}} = \sum_{a=1}^{A} \beta_a^{\top} h^a(Z,V) = \sum_{a,t,k} \beta_{a,k} \delta(z_t = k)\delta(v_t = a) \quad (4)$$

Notice Equation (3) assumes available labeled data V for the atomic actions present in each frame of D during training time. Nevertheless, similarly to Equation (2), it is possible to introduce latent variables and extend our model to the case that a dictionary of atomic actions needs to be learned.

At the third level of the hierarchy, we use the action vocabulary accumulated over all T frames to build a BoW representation for the underlying activity. Specifically, let $h^c(D)$ be the histogram corresponding to activity c in video D. Each entry a in $h^c(D)$ is given by:

$$h_a^c(D) = \sum_{t=1}^{T} \delta(v_t = a) \quad (5)$$

$$E_{\text{activity}} = \alpha_c^{\top} h^c(D) = \sum_{a=1}^{A} \sum_{t=1}^{T} \alpha_{c,a} \delta(v_t = a) \quad (6)$$

In terms of the energy terms associated to action and pose transitions in Equation (1), we depart from BoW representations and we introduce energy potentials that take into account temporal co-occurrences between poses and actions in neighboring frames. Specifically, let coefficients $\gamma_{a,a'} \in \mathbb{R}$ and $\eta_{k,k'} \in \mathbb{R}$ quantify co-ocurrence strength between neighboring pair of actions (a, a') or pair of poses (k, k'), respectively. Action and pose transitions energy potentials in Equation (1) are given by:

$$E_{\text{action transition}} = \sum_{a=1}^{A} \sum_{a'=1}^{A} \gamma_{a,a'} \sum_{t=1}^{T-1} \delta(v_t = a)\delta(v_{t+1} = a') \quad (7)$$

$$E_{\text{pose transition}} = \sum_{k=1}^{K} \sum_{k'=1}^{K} \eta_{k,k'} \sum_{t=1}^{T-1} \delta(z_t = k)\delta(z_{t+1} = k') \quad (8)$$

At the frame level, the previous model relies only on global image representations extracted at each frame. However, several works have shown the relevance of including local spatial information to boost recognition results [21]. Consequently, we account for local information by dividing each body pose at frame t into R spatial regions. Therefore, Equations (2), (4), (7), and (8) become, respectively:

$$E_{\text{pose}} = \sum_{r=1}^{R} \sum_{t=1}^{T} w_{z_{t,r},r}^{\top} x_{t,r} \quad (9)$$

$$E_{\text{action}} = \sum_{r=1}^{R} \sum_{a=1}^{A} \beta_{a,r}^{\top} h^{a,r}(Z,V) \quad (10)$$

$$E_{\text{action transition}} = \sum_{a,a',r} \gamma_{a,a',r} \sum_{t=1}^{T-1} \delta(v_{t,r} = a)\delta(v_{t+1,r} = a') \quad (11)$$

$$E_{\text{pose transition}} = \sum_{k,k',r} \eta_{k,k',r} \sum_{t=1}^{T-1} \delta(z_{t-1,r} = k)\delta(z_{t,r} = k') \quad (12)$$

3.3. Learning

We cast our formulation as an energy minimization problem. In particular, rather than first learning a dictionary of body poses and then learning classifiers for actions and activities, our goal is to learn all relevant parameters simultaneously using a multiclass max-margin approach. The input to our training algorithm is a set of M video sequences, where each video D_i containts annotations at the activity y_i and action V_i levels, and its set of T video frames is described by the set of feature vectors $X_i = (x_1, \dots, x_T)$. We aim to find optimal values for parameter sets α, β, W, γ , and η, as well as, slack variables ξ and latent variables Z, by solving the following max-margin learning problem:

$$\min_{\{\alpha,\beta,w,\gamma,\eta,\xi,Z\}} \frac{\lambda_1}{2RC}||\alpha||_F^2 + \frac{\lambda_2}{2RA}||\beta||_F^2 + \frac{\lambda_3}{2RK}||w||_F^2 + \frac{\lambda_4}{2R}||\gamma||_F^2 + \frac{\lambda_5}{2R}||\eta||_F^2 + \frac{\lambda_6}{M}\sum_{i=1}^{M} \xi_i \quad (13)$$

subject to the restrictions:

$$\max_{Z_i} E(X_i, Z_i, V_i, y_i) - E(X_i, Z, V, y) \geq \Delta((y_i, V_i), (y, V)) - \xi_i, \quad \forall y \in \mathcal{Y}, Z \in \mathcal{Z}, V \in \mathcal{V}, i \in [1...M] \quad (14)$$

A loss function that favors predicting the correct labels for each activity and set of atomic actions is given by:

$$\Delta((y_i, V_i), (y, V)) = \lambda_7 \delta(y \neq y_i) + \frac{\lambda_8}{RT} \sum_{r=1}^{R} \sum_{t=1}^{T} \delta(v_{t,r} \neq v_{(t,r)_i}) \quad (15)$$

By selecting a large value of λ_7, we give a large penalty when the activity is not predicted correctly. The second term adds a penalty proportional to the number of regions that are not labeled with the correct action according to V_i.

Given the loss function in Equation (15), the constrained optimization problem in Equation (13) is similar to a latent structural SVM case [39], therefore it can be solved using a Concave-Convex Procedure (CCCP) [40] which guarantees convergence to a local minimum. The CCCP algorithm alternates between maximizing Equation (13) with respect to the latent variables, and solving a structural SVM optimization problem [32] that treats latent variables as completely observed. Given space constraints we defer the details of this optimization to the supplementary material.

3.4. Inference

The input to the inference algorithm is a new video sequence with features X. The task is to infer the best activity label y^* and the best action labels V^*. Additionally, we also need to estimate latent variables Z.

$$y^*, V^*, Z^* = \underset{y,V,Z}{\operatorname{argmax}} E(X, Z, V, y) \quad (16)$$

We can solve this by exhaustively enumerating all values of y, and solving the following at each step:

$$V_y^*, Z_y^* = \underset{V,Z}{\operatorname{argmax}} E(X, Z, V, y) \quad (17)$$

Therefore, for each possible activity-class y, we must find V_y^* and Z_y^* frame-wise using:

$$v^*_{(t,r)_y}, z^*_{(t,r)_y} = \underset{v_t, z_t}{\operatorname{argmax}} \alpha_{y,r,v_t} + \beta_{v_t,r,z_t} + w^\top_{z_t,r} x_{t,r} + \gamma_{v_{t-1},v_t,r} + \eta_{z_{t-1},z_t,r} \quad (18)$$

See the supplementary material for further details.

4. Experiments

We validate our model with a series of evaluations using several activity datasets acquired with RGB-D cameras. We first report results on two previously released and publicly available datasets: MSR Action3D [22] and CAD120 [17]. Then, we use a new benchmark dataset with **Composable Activities** that we make available to the community.

Implementation details: In all experiments, we divide the body into $R = 4$ spatial regions: right arm, right leg, left arm, and left leg. Inspired by [10], at the lowest level, each body region r is represented by a feature vector x_r of 21 dimensions: 15 corresponding to line-to-line angles between body joints, and 6 corresponding to line-to-plane angles. In the case of MSR Action3D dataset that includes actions involving fine motions of hands and feet, we also incorporate to our descriptor the temporal derivative (velocity) of body joints associated to wrists and ankles.

To initialize latent variables Z, we obtain an initial dictionary of body poses by clustering low level descriptors X using the standard k-means algorithm. Using this initial dictionary, the value of each latent variable is obtained by associating the corresponding descriptor to its closest centroid. Afterwards, the pose dicctionary is estimated using a set of linear SVM classifiers, as in Eq. (2).

It is important to note that during training, our algorithm takes a global annotation for each video at the activity level, as well as per-frame annotations of the actions associated to each body region. At test time, these labels are not available and it is the task of our algorithm to infer them using the learned model. Furthermore, for training and evaluation purposes, we augment the annotations in each dataset with an additional *idle* or *background* action, at each frame where the subject is not executing any action.

4.1. MSR Action3D Dataset

The MSR Action 3D dataset consists of 7 subjects performing 20 possible actions. For more details about this dataset please refer to [22]. In our experiments, we use 552 sequences, and we test our model using cross-subject validation. We omit some videos as they have missing joints in the last half of the sequence. This dataset only includes annotations at a single complexity level, and each video is associated with a single global action label. In order to better showcase the hierarchical capabilities of our model, we keep the global activity label but also annotate all frames with the given action class, except those frames where the subject is standing still, which we label as *background*.

Our model achieves an action classification accuracy of 89.5%, a recognition performance that is on par with the state-of-the-art. Although this dataset does not provide a rich hierarchy of complex activities composed by atomic actions, this result allow us to validate that our model performs well on the task of single action recognition.

As in the case of alternative techniques, most of the actions can be recognized with almost perfect accuracy, but some actions are still dificult to discriminate, such as *hand catch* and *high throw*, due to highly similar movements. Table 1 shows the accuracy of our method in comparison to state-of-the-art approaches.

4.2. CAD120 Dataset

The CAD120 dataset is introduced in [17]. It is composed of 124 videos that contain activities in 10 clases per-

Algorithm	Accuracy
Our method	89.46%
J. Wang et al. [34]	88.20%
C. Wang et al.[33]	90.22%
Oreifej and Liu [26]	88.89%
Xia and Aggarwal [37]	89.30%

Table 1. Recognition accuracy of our method compared to state-of-the-art methods using MSR Action3D dataset.

Algorithm	Average precision	Average recall
Our method	32.6%	34.58%
[17]	27.4%	31.2%
[30]	23.7%	23.7%

Table 2. Recognition accuracy of our method compared to state-of-the-art methods using CAD120 dataset.

formed by 4 actors. Activities are related to daily living: *making cereal*, *stacking objects*, or *taking a meal*. Each activity is composed of simpler actions like *reaching*, *moving*, or *eating*. In this database, human-object interactions are an important cue to identify the actions, so object locations and object affordances are provided as annotations. Performance evaluation is made through leave-one-subject-out cross-validation. Given that our method does not consider objects, we use only the data corresponding to 3D joints of the skeletons. As shown in Table 2, our method outperforms the results reported in [17] using the same experimental setup. It is clear that using only 3D joints is not enough to characterize each action or activity in this dataset. As part of our future work, we expect that adding information related to objects will further improve accuracy.

4.3. A New Dataset: Composable Activities

We introduce a new benchmark dataset, **Composable Activities**, consisting of 693 videos that contain activities in 16 classes performed by 14 actors. We capture RGB-D data for each sequence using a Microsoft Kinect sensor and estimate position of relevant body joints using [24]. We only used body joint positions from the estimated skeleton in each frame to compute our descriptors. Each activity in this dataset is spatio-temporally composed by a number of mid-level (atomic) actions. The total number of actions in the videos is 26, while the number of actions that compose each particular activity fluctuates between 3 to 11 actions. For instance, the activity *walk while hand waving* has a spatio-temporal composition of 3 single actions: *walk*, *hand wave*, and *idle*; while the activity *composed-activity-4* is composed of 11 single actions: *idle*, *walk*, *call a friend with hands*, *hand wave*, *talking on cellphone*, *pick from the floor*, *dial cellphone*, *put an object*, *pick cellphone from pocket*, and *put cellphone in pocket* (see Figure 1). The skeleton data and annotations can be downloaded our project webpage, http://web.ing.puc.cl/ialillo/ActionsCVPR2014.

In our leave-one-subject-out experimental setup, the accuracy of our model is 85.7%, when using $K = 50$ poses for each body part (a total of 200 poses), which provides a good compromise between model complexity and accuracy. We also set the model parameters λ_7 to 500, λ_8 to 100, and λ_9 to 20. In general, we use cross-validation to adjust the value of all our main parameters.

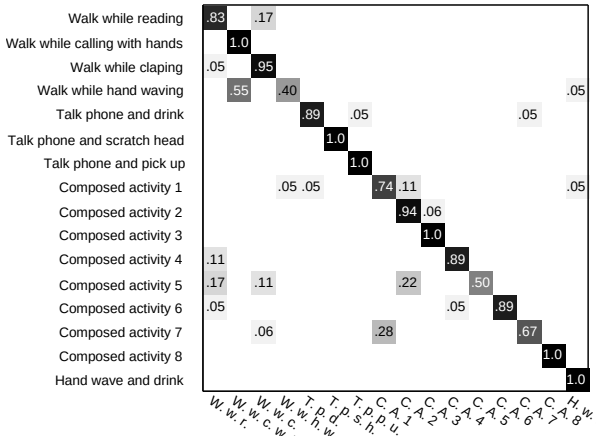


Figure 3. Confusion matrix for the activity classification task in the new Composable Activities dataset.

We compare the performance of our method with respect to three baselines techniques: a BoW representation plus a lineal SVM classifier (BoW-approach), a version of our model without learning the pose dictionary (H-BoW-approach), and a Hidden Markov Model approach (HMM-approach). In the case of BoW-approach we use k-means to obtain a pose dictionary that is used to quantize the observed poses. We build a global histogram of poses using all frames of the sequence and enhance it with a version of Spatial Pyramids (SPM). The accuracy of this baseline is 67.2%. Our model demonstrates a substantial accuracy improvement, exploiting the ability to model activities and actions, and jointly learning a pose dictionary. A second baseline consists of a simplified version of our hierarchical model that does not learn the pose dictionary, but uses a fixed pose quantization given by k-means. In this case, the accuracy drops by 11%, which supports the inclusion of our discriminative learning scheme to learn the pose dictionary. The third baseline is an HMM model, which is learned using atomic actions as states and poses as observed variables. A model is learned independently for each class. At test time, we score a new sequence using all models, and select the activity label that corresponds to the model with highest log-likelihood. This baseline obtains an accuracy of 76.5%. Recognition rates are summarized in Table 3.

Effects of size of pose dictionary: Our method is relatively robust to the size of the pose dictionary. A low number of poses per body part (5 to 20) lacks reprensentativity, and a high number increases the computational load. In our experiments, we observe similar performance for the case of 50, 100 and 150 poses per body part. When testing with

Algorithm	Codebook size	Accuracy
Our method	200 (learned)	**85.7%**
Our method	600 (learned)	**82.9%**
BoW	200 (fixed)	67.2%
BoW	600 (fixed)	62.3%
H-BoW	200 (fixed)	74.2%
H-BoW	600 (fixed)	71.5%
HMM	200 (fixed)	76.5%
HMM	600 (fixed)	72.3%

Table 3. Recognition accuracy of our method compared to three baselines: Bag-of-Visual-Features (BoW), our method but without learning pose dictionary (H-BoW), and a Hidden Markov Model approach (HMM).

25 poses the accuracy drops by 6%. We did not test larger dictionaries due to the processing time, which is quadratic with respect to dictionary size. We chose 50 poses per body part in our experiments as a compromise between good accuracy and processing speed.

Importance of transition terms in the model: When we simplify our model by fixing the pose dictionary and dropping the energy terms related to action and pose transitions, we observe a drop in accuracy of 11.2%. If we learn the pose dictionary, but ignore the temporal transition components γ and η, accuracy drops by 4.8%. As expected, learning temporal cooccurrence improves the accuracy of our method, as it links poses and actions over time. In our current model, we use a single frame correlation; this short-term relation could be expanded to middle or long term correlations, with the cost of an increased running time.

Action annotation: The hierarchical structure and compositional properties of our model enable it to perform per-frame annotation of the atomic actions that compose each activity, and to indicate which body parts are associated to the atomic actions present in a frame, as well as the temporal span of each action. We illustrate this capability in Fig.4. The accuracy of the mid-level action prediction can be evaluated as in [36]. We first get segments of the same predicted action in each sequence, and then compare these segments with ground truth action labels from annotated data. The inferred label of the segment is assumed correct if a detected segment is completely contained in a ground truth segment with the same label, or if the Jaccard Index of the segment and the ground truth of the same action label is greater than 0.6. Using these criteria, the accuracy of the mid-level actions is 70.2%. In many of the mistakes, the wrong action label only affects a single part, and the model is still able to correctly predict the activity label of the sequence.

Robustness to occlusion: Our method is also capable of inferring action and activity labels even if some joints are not observed. To illustrate this, we simulate an occluded part by fixing it to the position observed in the first frame. We select a part to be occluded in every sequence using a uniform sampling. In this scenario, the accuracy of our

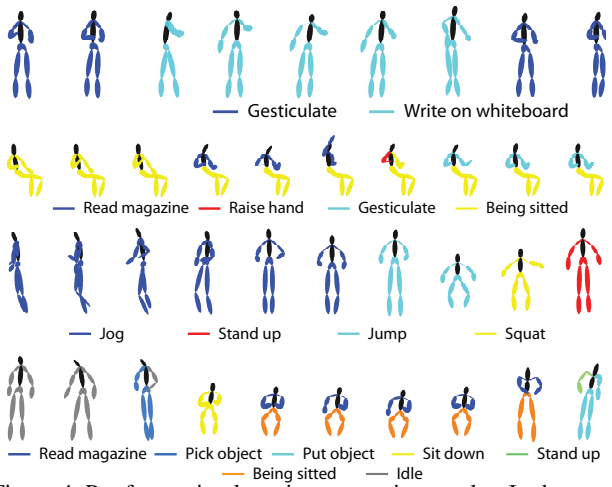


Figure 4. Per-frame simple action annotation results. In these examples, our algorithm correctly classifies the overall activity category. Furthermore, it is able to correctly predict the atomic actions that compose each activity and which body parts contribute to those actions. Each body part is colored according to the predicted action label.

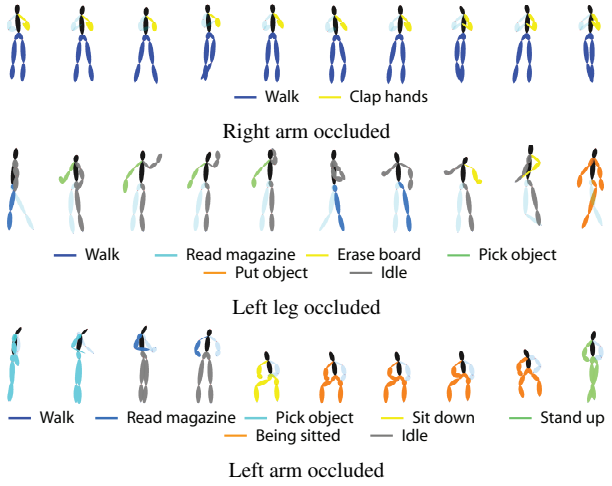


Figure 5. The occluded body parts are depicted in light blue. When an arm or leg is occluded, our method still provides a good estimation of actions in each frame.

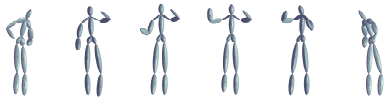
model drops by 7.2%, while the drops in performance of BoW is (12.5%) and HMM (10.3%). Also, Fig. 5 shows some qualitative results.

5. Conclusions and Future Work

We present a novel hierarchical compositional model to recognize human activities using RGB-D data. The proposed method is able to jointly learn suitable representations at different abstraction levels leading to compact and robust models, as shown by the experimental results. In particular, our model achieves powerful multi-class discrimination while providing useful annotations at the intermediate semantic level. The compositional capabilities of our model also bring robustness to partial body occlusions.

Ground truth: Walk while calling with hands

Prediction: Walk while hand waving



Ground truth: Composed activity 1

Prediction: Talk phone and drink

Ground truth: Hand wave and drink

Prediction: Talk phone and scratch head

Figure 6. Failure cases. Our algorithm tends to confuse activities that share very similar body postures.

Acknowledgements This work was funded by FONDECYT grant 1120720, from CONICYT, Government of Chile, and LACCIR grant RFP1212LAC005. I.L. is supported by a PhD studentship from CONICYT. J.C.N. is supported by a Microsoft Research Faculty Fellowship.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis. *ACM Computing Surveys*, 43(3):1–43, Apr. 2011.

[2] M. R. Amer and S. Todorovic. Sum-product networks for modeling activities with stochastic structure. In *CVPR*, 2012.

[3] Y. Bengio. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[4] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[5] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *TPAMI*, 23(3):257–267, 2001.

[6] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[7] W. Brendel and S. Todorovic. Activities as Time Series of Human Postures. In *ECCV*, pages 721–734, 2010.

[8] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, pages 778–785. IEEE, Nov. 2011.

[9] A. Castrodad and G. Sapiro. Sparse modeling of human actions from motion imagery. *IJCV*, 100:1–15, 2012.

[10] C. Chen, Y. Zhuang, F. Nie, Y. Yang, F. Wu, and J. Xiao. Learning a 3D Human Pose Distance Metric from Geometric Pose Descriptor. *IEEE TVCG*, 17(11):1676–1689, Dec. 2010.

[11] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[12] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. In *VSPETS*, 2005.

[13] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[14] A. Gaidon, Z. Harchaoui, and C. Schmid. Temporal Localization of Actions with Actoms. *IEEE TPAMI*, Mar. 2013.

[15] N. Ikizler and D. A. Forsyth. Searching for Complex Human Activities with No Visual Examples. *IJCV*, 80(3):337–357, 2008.

[16] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.

[17] H. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from RGB-D videos. *IJRR*, 32(8):951–970, 2013.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[19] I. Laptev. On Space-Time Interest Points. *IJCV*, 64(2-3):107–123, 2005.

[20] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[22] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. *CVPR*, 2010.

[23] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *CVPR*, 2008.

[24] Microsoft. Kinect for Windows SDK, 2012.

[25] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In *ECCV*, 2010.

[26] O. Oreifej and Z. Liu. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In *CVPR*, 2013.

[27] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *NIPS*, 2003.

[28] S. Savarese, J. Winn, and A. Criminisi. Discriminative Object Class Models of Appearance and Shape by Correlatons. In *CVPR*, 2006.

[29] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011.

[30] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured Human Activity Detection from RGBD Images. In *ICRA*, 2012.

[31] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008.

[32] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

[33] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. *CVPR*, 2013.

[34] Y. Wang and D. Forsyth. Discriminative Hierarchical Part-based Models for Human Parsing and Action Recognition. *JMLR*, 13:3075–3102, 2012.

[35] Y. Wang and G. Mori. Hidden part models for human action recognition: Probabilistic versus max margin. *PAMI*, 33(7):1310–1323, 2011.

[36] P. Wei, N. Zheng, Y. Zhao, and S.-C. Zhu. Concurrent Action Detection with Structural Prediction. *ICCV*, pages 3136–3143, 2013.

[37] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, 2013.

[38] L. Xia, C.-C. Chen, and J. K. Aggarwal. View Invariant Human Action Recognition Using Histograms of 3D Joints. In *CVPR Workshop*, 2012.

[39] C. Yu and T. Joachims. Learning structural SVMs with latent variables. In *ICML*, 2009.

[40] A. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.