THE ROMANIAN ACADEMY

SIMION STOILOW INSTITUTE OF MATHEMATICS

Models for automatic recognition and detection of human activities in video

Author,
Mihai ZANFIR

Supervisor,
C.S. I Dr. Cristian SMINCHIȘESCU

PH. D. THESIS SUMMARY

Bucharest
2018

Chapter 1 - Introduction. This thesis focuses on the problem of human activity recognition based on three-dimensional representations. This is an important problem, with numerous applications in domains such as smart video surveillance systems, entertainment, video retrieval, autonomous driving and human-robot interaction.

The difficulty of human sensing stems from the challenges of human body part detection and from the coupling of the recognition and reconstruction tasks. For the former, one needs to take into consideration the variability of human pose and shape, the scene diversity and the background clutter. For the latter, one has to account for the inter- and intra- class variability of different action types, and account for the prediction errors of the 3d human pose reconstructions models used.

Chapter 2. We propose a fast, non-parametric framework for low-latency human action and activity recognition. We illustrate the model for the 3d pose reconstructions from depth sensors and show how this framework naturally supports low-latency recognition, one-shot learning, and action detection in unsegmented video data, with high accuracy. Central to our methodology is the Moving Pose descriptor–a novel frame-based dynamic representation that captures not only the 3D body pose, but also differential properties like the speed and acceleration of the human body joints within a short time window around the current frame. We argue that due to physical constraints like inertia, or latency in muscle actuation, the body movements associated with an action can often be well approximated by a quadratic function, expressed in terms of the first and second derivatives of the body pose with respect to time.

Inspired by this, for each frame of a video sequence we compute the Moving Pose Descriptor (MP), as a concatenation of the normalized 3D pose $\mathbf{P} = [\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n}]$ and its first and second order derivatives $\delta\mathbf{P}(t_0)$ and $\delta^2\mathbf{P}(t_0)$. The derivatives are estimated numerically by using a temporal window of 5 frames centered at the current one processed: $\delta\mathbf{P}(t_0) \approx \mathbf{P}(t_1) - \mathbf{P}(t_{-1})$ and $\delta^2\mathbf{P}(t_0) \approx \mathbf{P}(t_2) + \mathbf{P}(t_{-2}) - 2\mathbf{P}(t_0)$. For better numerical approximation we first smooth each coordinate of the normalized pose vector, along the time dimension, with a 5 by 1 Gaussian filter ($\sigma = 1$).

The proposed MP descriptor encodes pose and kinematic information to describe action segments. In order to emphasize its discriminative power and for training flexibility

(including one-shot learning) we use a non-parametric action classification scheme based on k-nearest-neighbors (kNN). Our basic low-latency classification method for a test sequence is as follows: at time t, after observing t test frames, let each of their kNN descriptors from the training pool vote for its class, for a total of kt votes. For decision, we apply a simple rejection scheme. If the accumulated vote of the most supported class c_j is high enough compared to the other classes, and enough frames have been observed, we report the class with the largest number of votes, *i.e.* output c_j if $\max_j s(c_j, t) \geq \theta$, with s an additive voting model and θ a confidence threshold. We learn the value of the vote for each frame in the training set, such that the more representative frames of a certain action are assigned a higher discriminative power. Our full approach is to incorporate global temporal information within a kNN framework to gate the search for nearest neighbors only to samples that are located at a similar position in the training sequence, with respect to the first frame.

By combining discriminative local moving pose descriptors like MP with a temporal aware classification scheme, we can now account for two important aspects in action classification: the discriminative power of key poses as well as their local dynamics, and the global temporal course of an action.

Tabela 1: Recognition comparison on the MSR Action3D dataset.

Method	Accuracy(%)
Recurrent Neural Network [10]	42.5
Dynamic Temporal Warping [12]	54
Hidden Markov Model [9]	63
Latent-Dynamic CRF [11]	64.8
Canonical Poses [2]	65.7
Action Graph on Bag of 3D Points [7]	74.7
Latent-Dynamic CRF [11] + MP	74.9
EigenJoints [18]	81.4
Actionlet Ensemble [16]	88.2
MP (Ours)	**91.7**

Our system (see Table 1) improves over the current state-of-the-art by 3.5% on the MSR Action3D dataset. This dataset consists of temporally segmented action sequences captured by an RGB-D camera. The 3D skeleton, represented as a set of 3D body joint positions, is

available for each frame, being tracked with the method of [3].

Chapter 3. We further develop representations not only based on human pose and we generalize the Moving Pose (MP) formulation [20] and consider pairwise topological relationships between simple features (e.g. 2D or 3D positions). We also exploit the kinematic properties of these relations, by their first and second-order derivatives w.r.t time, to form *spatiotemporal* sub-graphs as frame descriptors suitable for recognizing different types of actions. Unlike the Moving Pose, these pairwise relations are formed at a higher level of abstraction. Instead of measuring exact, real-valued geometry, we construct models based on soft classifiers that respond to relationships between different human joints and objects. We consider three types of topological relations: *top-down*, *left-right* and *front-back*. Let us look at the 2D case, the same formulation being immediately extended to 3D. Given a pair of feature points (i, j), with locations $p_i = (x_i, y_i)$ and $p_j = (x_j, y_j)$, there are two types of topological relationships modeled with a categorical predictor using logistic functions, $R_x(x_i, x_j)$ and $R_y(y_i, y_j)$. R_x and R_y are soft binary classifiers that respond to *left-right* and *top-down* topological relationships:

$$R_x(x_i, x_j) = \frac{1}{1 + \exp(-w_x(x_i - x_j))} \quad R_y(y_i, y_j) = \frac{1}{1 + \exp(-w_y(y_i - y_j))}. \qquad (1)$$

Given the locations of human body joints and objects in the scene, there are exponentially many possible subsets of relations to consider. It turns out that only a small group of relations are effective for classification. Finding this small set requires an efficient search procedure in a very large space of possible sets. For example, given 20 body joints, the number of possible relations (all three types) in our case is 570, so there are 2^{570} possible MR frame descriptors (or sets of relations). Let D be the set of all possible MR's, of size 2^{570}. We formulate the task of finding the optimal descriptor d_a^* for a given class a as maximizing the empirical expected difference between the soft classification response on positive sequences and the response on negative ones:

$$d_a^* = \underset{d \in D}{\operatorname{argmax}} \left(\frac{1}{N_a} \sum_{s \in S(a)} C(s,d) - \frac{1}{N_{\neg a}} \sum_{s \in S(\neg a)} C(s,d) \right). \quad (2)$$

Our proposed approach is to start with a stochastic method, extended from [6], to estimate the relevance of each relation and joint separately, and then form, in a greedy manner, an initial Moving Relations descriptor, with good performance on all action classes. Starting from this common MR d_0, we then follow an evolutionary search strategy, by adapting the Genetic Algorithm (GA) proposed in [5], to learn different descriptors d_a^* that are optimized for each action class.

Tabela 2: Recognition comparison on the CAD-120 dataset.

Method	Accuracy(%)
Moving Pose [20]	67.5
Koppula et. al. [4]	83.1
Discovered Moving Relations without GA (Ours)	93.3
Discovered Moving Relations with GA (Ours)	**99.2**

Our experiments demonstrate the power of MR, which in combination with a modified kNN classification scheme, significantly outperforms more sophisticated current methods. Unlike most methods, ours is robust to missing features and applicable to such situations with no modification.

Chapter 4. In this thesis chapter, we propose a deep multitask architecture for fully automatic 2d and 3d human sensing, including recognition and reconstruction, in monocular images. The system predicts the figure-ground segmentation, body-part labelling at pixel level, and estimates the 2d and and 3d pose of the person in the scene. Conceptually, each of our stages of processing produces recognition and reconstruction estimates and is constrained by specific training losses. Each task consists of a total of six recurrent stages which take as input the image, the results of previous stages of the same type (except for the first one), as well as inputs from other stages (2d pose estimation feeding into semantic body part segmentation, and both feeding into 3d pose reconstruction). The inputs to each stage are individually processed and fused via convolutional networks in order to

produce the corresponding outputs. The 2d pose estimation task is based on a recurrent convolutional architecture similar to [17]. Given an RGB image $I \in \mathbb{R}^{w \times h \times 3}$, we seek to correctly predict the locations of N_J anatomically defined human body joints $p_k \in \mathcal{Z} \subset \mathbb{R}^2$, with $k \in \{1 \dots N_J\}$.

For semantic body part segmentation (body part labeling) we assign each image location $(u, v) \in \mathcal{Z} \subset \mathbb{R}^2$ one of N_B anatomical body part labels (including an additional label for background), b_l, where $l \in \{1 \dots N_B\}$. At each stage t, the network predicts, for each pixel location, the presence probability of each body part, $B^t \in \mathbb{R}^{w \times h \times N_B}$. During the first stage of processing, we use convolutional representations based on the image and the 2d pose belief maps J^1 in order to predict the current body labels B^1. For each of the following stages, we also use the information present in the body labels at the previous stage, B^{t-1}, and rely on a series of four convolutional layers c_B^t that learn to combine inputs obtained by stacking image features $\mathbf{x}$ and B^{t-1}.

The 3d reconstruction module leverages information provided by the 2d joint and body part labeling feature maps J^t and B^t. Additionally, we insert a trainable function c_D^t, defined similarly to c_B^t, over image features, in order to obtain body reconstruction feature maps D^t. The module follows a similar flow as the previous ones: it reuses estimates at earlier processing stages, R^{t-1}, together with S^t and D^t, in order to predict the reconstruction feature maps R^t. The processing stages and dependencies of this module are shown in fig. 0-1.

The design allows us to tie a complete training protocol, by taking advantage of multiple datasets that would otherwise restrictively cover only some of the model components: complex 2d image data with no body part labeling and without associated 3d ground truth, or complex 3d data with limited 2d background variability. In detailed experiments based on several challenging 2d and 3d datasets (LSP, HumanEva, Human3.6M), we evaluate the sub-structures of the model, the effect of various types of training data in the multitask loss, and demonstrate that state-of-the-art results can be achieved at all processing levels.

Chapter 5. Finally, we introduce fine-grained action and emotion recognition tasks defined on non-staged videos, recorded during therapy sessions of children with autism from the multi-modal DE-ENIGMA [14] dataset. The sessions are either therapist-only or robot-

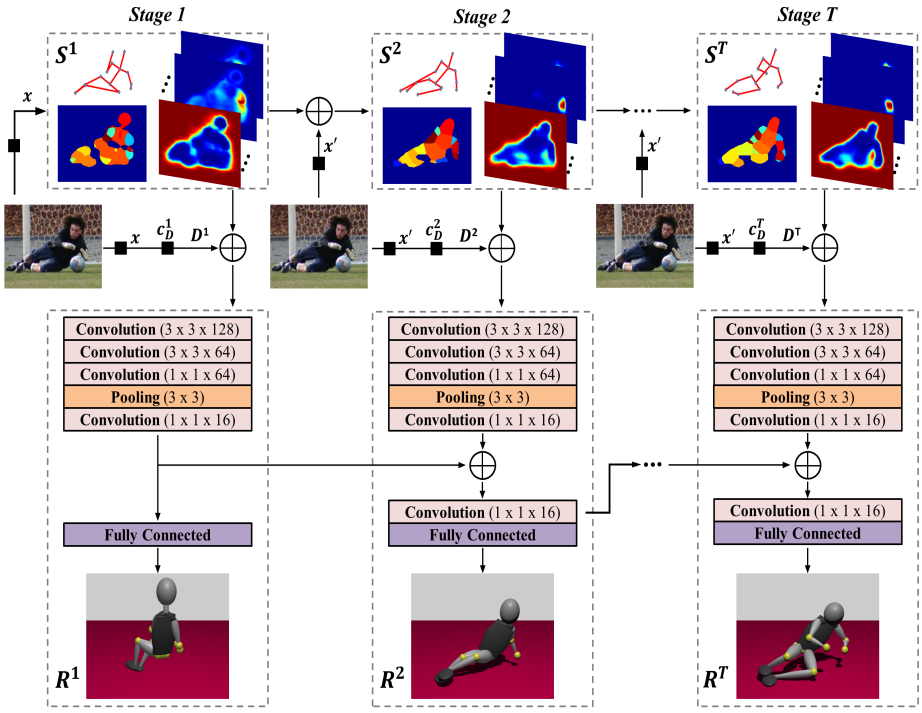

Figura 0-1: Our multitask multistage 3d reconstruction module R^t, combines 3d processing with information from semantic modules, S^t.

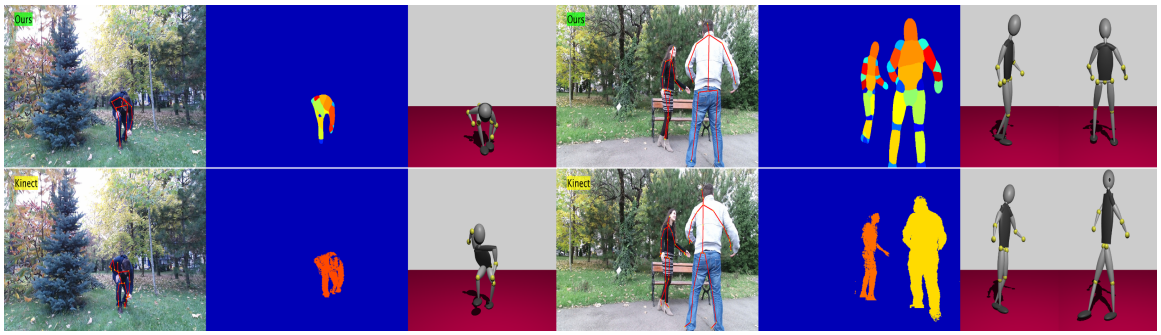

Figura 0-2: Qualitative comparisons for segmentation and reconstruction between our RGB model (top row) and the ones of a commercial RGB-D Kinect for Xbox One system (bottom row). Our model produces accurate figure-ground segmentations, body part labeling, and 3d reconstruction for some challenging poses.

assisted; the former are captured for control purposes, while the latter are those of interest for this chapter. In robot-assisted sessions a child and a therapist sit in front of a table on which a robot is placed. The therapist remotely controls the robot and uses it to engage the child in the process of learning emotions. The sessions consist of a 'free-play' part (where the child plays with toys of his choice), and an actual therapy part. The therapy is based on scenarios in which the therapist shows cards depicting various emotions (happy, sad, angry, etc.) which are also reproduced by the robot, and the child must match the emotions to those performed.

The therapy scenarios cover a wide variety of body gestures and actions performed by children. We have annotated a total of 3757 sequences, with an average duration of 2.1 seconds. The annotation of therapy videos relies on an extensive web-based tool developed by us that can (i) select temporal extents and (ii) assign them a class label. Features that improve the annotation experience such as shortcuts for precise temporal adjustments, current selection replays, previous annotations filtering and visualization, or user session management, are also included.

The experiments presented in this chapter use a subset of 2031 annotated sequences spanning over 24 classes common to all children. Even if the selected classes refer to children behavior, some of them relate to the therapist, e.g., *Pointing to therapist*, *Turning towards therapist*. We refer to those as interacting sequences. Among the annotated sequences, around a third (749 out of $2,031$) are interacting sequences.

Our long term goal is to automatically interpret and react to a child's actions in the challenging setting of a therapy session. In order to understand the child, we rely on high-level features associated to her/his 3d pose and shape. We use the previously introduced DMHS and improve it on the 3d pose estimation taks of partially visible people (*i.e.* DMHSPV).

We rely on a feedforward-feedback model presented in the paper [19] to combine human detection, 2d and 3d pose prediction from DMHSPV with a shape-based volumetric refinement based on a SMPL body representation [8] – DMHS-SMPL-F and the temporal smoothed variant DMHS-SMPL-T.

We experiment with several skeleton-based action recognition models and perform ablation studies with different types of 2d and 3d human body reconstructions. We use a

cross-validation setting on children where we consider only the upper-body joints of the human skeleton.

Pose Feature	MP - Child	MP - Child + Therapist
Kinect [15]	46.96%	**47.49%**
DMHSPV	32.92%	34.95%
2D [1]	40.83%	44.14
DMHS-SMPL-F	43.53%	45.07%
DMHS-SMPL-T	44.20%	45.68%

Tabela 3: Comparative results for different pose estimation methods for action classification when using the Moving Pose framework. We also investigate the impact of modeling the therapist in the classification accuracy.

A video selection from [14], including those 7 children used for action classification experiments, was also annotated with continuous emotions in a valence-arousal space by 5 specialized therapists. We pre-process the data as in [13] to obtain per frame values for each annotator and align them to obtain a reliable ground-truth valence/arousal signal.

Emotion Axis	Pose Feature	RMSE ↓	PCC ↑	SAGR ↑
Valence	Kinect	0.116	**0.184**	0.787
	DMHS-SMPL-T	**0.099**	0.169	**0.844**
Arousal	Kinect	0.111	0.345	0.973
	DMHS-SMPL-T	**0.107**	**0.388**	**0.977**

Tabela 4: Continuous emotion prediction. Using 3d skeleton estimates of DMHS-SMPL-T, we obtain better or similar results compared to the 3d skeleton produced by Kinect.

Bibliografie

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[2] C. Ellis, S.Z. Masood, M.F. Tappen, J.J. LaViola Jr., and R. Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *IJCV*, August 2012.

[3] J. Shotton et al. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011.

[4] H. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *ICML*, 2013.

[5] R. Leardi, R. Boggia, and M. Terrile. Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, 6, 1992.

[6] S. Li, E.J. Harner, and D.A. Adjeroh. Random knn feature selection-a fast and stable alternative to random forests. *BMC bioinformatics*, 12(1), 2011.

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

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH*, 34(6):248:1–16, 2015.

[9] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *ECCV*, 2006.

[10] James Martens and Ilya Sutskever. Learning recurrent neural networks with hessian-free optimization. In *ICML*, 2011.

[11] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR*. IEEE Computer Society, 2007.

[12] Meinard Müller and Tido Röder. Motion templates for automatic classification and retrieval of motion capture data. In *SCA*, 2006.

[13] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In *Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*. German Research Center for AI (DFKI), 2010.

[14] J. Shen, E. Ainger, A. M. Alcorn, S. Babović Dimitrijevic, A. Baird, P. Chevalier, N. Cummins, J. J. Li, E. Marchi, E. Marinoiu, V. Olaru, M. Pantic, E. Pellicano, S. Petrovic, V. Petrovic, B. R. Schadenberg, B. Schuller, S. Skendžić, C. Sminchisescu, T. T. Tavassoli, L. Tran, B. Vlasenko, M. Zanfir, V. Evers, and Consortium De-Enigma. Autism data goes big: A publicly-accessible multi-modal database of child interactions for behavioural and machine learning research. *International Society for Autism Research Annual Meeting*, 2018.

[15] 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 single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304. IEEE Computer Society, 2011.

[16] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[17] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, June 2016.

[18] X. Yang and Y. Tian. Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor. In *CVPR Workshops*, 2012.

[19] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes – The Importance of Multiple Scene Constraints. In *CVPR*, 2018.

[20] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The "Moving Pose": An Efficient 3D Kinematics Descriptor for Low-Latency Action Detection and Recognition. In *International Conference on Computer Vision*, December 2013.