Latent Spatio-temporal Models for Action Localization and Recognition in Nursing Home Surveillance Video

Yuke Zhu, Tian Lan, Yijian Yang, Steven N. Robinovitch, Greg Mori
Simon Fraser University
Burnaby, BC, Canada
yukez@sfu.ca, tla58@sfu.ca, yijiany@sfu.ca, stever@sfu.ca, mori@cs.sfu.ca

Abstract

This paper presents an application of vision-based monitoring of long-term care facility residents. We develop an algorithm to detect events of interest, particularly falls by elderly residents. The algorithm uses a max-margin latent variable approach with spatio-temporal locations of the person in the video as latent variables. The recently developed Action Bank descriptor is utilized as a rich feature representation for each frame. Empirical results demonstrate the effectiveness of this method.

1 Introduction

In this paper we present an algorithm for action recognition in surveillance videos. We focus on an application in monitoring residents in long-term care facilities – detecting actions, particularly falls by residents.

There are myriad potential uses for a robust system for such monitoring. Among them is developing interventions to reduce the number and serverity of falls by long-term care facility residents. A video-based system that automatically detects falls should allow for more prompt medical response from care providers. Furthermore, such a system could provide objective data on the causes and circumstances of falls, which are currently lacking [9]. If objective data on the frequencies of different types of fall events can be gathered, intervention strategies can be prioritized, and cost-benefit analyses can be conducted.

Action recognition in real-world surveillance video is a challenging problem. Modeling the spatio-temporal location of the person in the video builds a figure-centric representation that focuses recognition on the action of the person, and can be robust to cluttered scenes and variablility in person position. In this paper we develop a latent variable framework that encodes this information. Rather than running a processing pipeline that includes generic human detection and tracking, we treat the location of the human performing the action as a latent variable, and infer its location automatically. We do not require a human detector to initialize the inference process, and utilize a state-of-the-art template based action representation, Action Bank [10] to describe regions of video. We demonstrate empirically that this approach is effective.

2 Previous Work

Vision-based action recognition is an active area of research. Weinland et al. [12] provide a comprehensive recent survey. Below, we briefly review recent work.

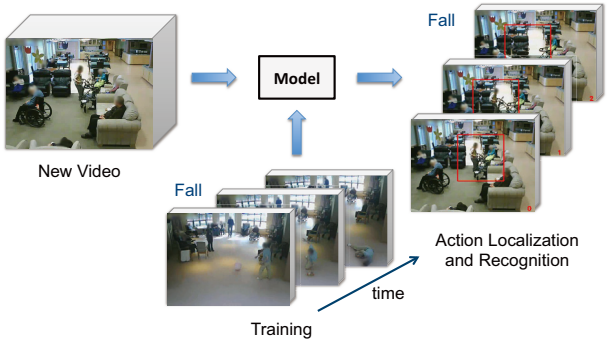

Figure 1. An overview of the proposed model. We train our model on weakly labeled video data with an action label for each clip. Given a new video, we use the model to localize and recognize an action in the video. Faces blurred for privacy.

Hidden Markov models (HMMs) have been widely used in action recognition, from early work [13] through to recent latent sequence models [11]. Our work uses a discriminative HMM-like structure which includes the position of the person as a latent variable.

A variety of methods have been developed for action recognition that utilize a detection and tracking framework that is followed by subsequent action classification. Choi and Savarese [3] developed a novel variant of this approach, jointly tracking and performing action recognition for groups of people together. Huang et al. [6] also treat tracking as latent variable for action recognition, though of individuals, and starts from individual tracklets. Both of these works leverage human detections and are applied to scenes with limited variability in human pose.

Lan et al. [8] develop a method for latent localization of action in more diverse videos (UCF-Sports), using a low-threshold human detection to guide inference. Bilen et al. [2] localize actions as latent sub-volumes in videos, using a bag-of-words representation. Yao et al. [14] devise a hough transform voting scheme for action recognition that does not require explicit detection a priori. Amer et al. [1] develop a framework for multi-scale analysis of human activities in an AND-OR graph formalism. A bottom-up cost sensitive inference procedure is used to detect low-level actions.

The method we develop follows in the spirit of latent localization methods [8, 2], though uses a more descriptive feature representation and more flexible latent variable structure that does not require a priori human detection.

3 Latent Spatio-temporal Model

In this paper, we propose a discriminative latent spatio-temporal model for action localization and recognition in surveillance video. We start by introducing our representation for videos, then give a detailed mathematical formulation of our model.

Intuitively, a complex action can be decomposed into a sequence of simpler atomic actions. For instance, a falling action can be divided into three states including losing balance, descending and lying down. There are strong temporal correlations between pairs of states, e.g. losing balance should happen before descending which is followed by lying down. We use a HMM-like representation to capture such intuition, where the temporal links between different states is encoded in a chain structure.

In order to perform action localization, we introduce a latent variable for each temporal segment of a video. The latent variable encodes where in space-time an action is occuring. In practice, the transitions from one state to another are constrained by a distance threshold enforcing that the spatio-temporal volumes on the action of interest should change smoothly over time.

3.1 Video Representation

We define two parameters in our model, a temporal parameter T and a spatial parameter S, for each dataset. We uniformly divide each video into T temporal segments. We further split each temporal segment into S spatial regions. Hence, a video sequence is partitioned into $T \times S$ spatio-temporal volumes. We use the recently proposed Action Bank descriptor [10] to represent each spatio-temporal volume. Action Bank is a high-level video representation for action recognition. It contains a collection of 205 individual action detectors, and the standard descriptor contains a concatenation of volumetric max-pooled detection volume features from each detector.

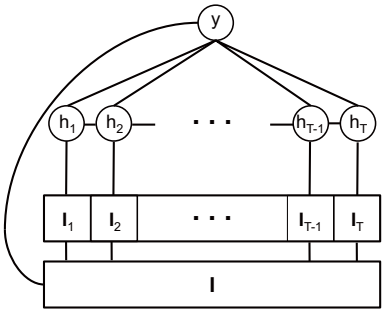

Figure 2. Illustration of our model. $\mathbf{I}$ denotes the input video. I_i denotes the i-th temporal segment. h_i denotes the spatial location of the discriminative latent sub-volume in I_i. y is the action label of the video $\mathbf{I}$.

3.2 Model Formulation

An action label $y \in \mathcal{Y}$ is assigned to each video, where $\mathcal{Y}$ is the domain of action labels. Each training video $\mathbf{I}$ is divided into T temporal segments, $\mathbf{I} = (I_1, I_2, \ldots, I_T)$, where I_i denotes the i-th temporal segment of the video. Each segment I_i is further split into S equal-sized spatio-temporal volumes, denoted as $\mathcal{R}_i = \{r_{i1}, r_{i2}, \ldots, r_{iS}\}$, where r_{ij} denotes the j-th spatio-temporal volume in the i-th segment. We use $\phi(I_i; r_{ij})$ to represent the Action Bank features extracted from the spatio-temporal volume r_{ij}.

We encode the latent spatial location of the person in the video as a vector $\mathbf{h} = (h_1, h_2, \ldots, h_T)$, where $h_i \in \mathcal{R}_i$ denotes the spatio-temporal volume containing the person performing the action in the i-th temporal segment. The feature vector for the entire video $\Phi(\mathbf{I}; \mathbf{h})$ is defined as the concatenation of $\phi(I_i; h_i)$, $1 \leq i \leq T$. In our model, $\mathbf{h}$ is treated as latent variables to be inferred simultaneously with action recognition.

We encode temporal smoothness into our model by setting a distance threshold D to the transitions between adjacent states h_i and h_{i+1}. The distance is measured by the Euclidean distance from the centre of the i-th spatio-temporal volume to the i+1-th volume. Transitions can only be made between temporal segments closer than the threshold D. This constraint reduces the domain $\mathcal{H}$ of all possible latent variables. A graphical illustration of our model is shown in Fig. 2.

A training example is represented as a tuple $\{\mathbf{I}, y\}$, where $\mathbf{I}$ is video itself and y is the action label assigned to the video. Inspired by the latent SVM [15, 5], we use the discriminative scoring function $f_\omega(\mathbf{I})$ to model the dependencies among the variables, where $\boldsymbol{\omega}$ is a model parameter to be optimized:

$$f_\omega(\mathbf{I}) = \max_{\mathbf{h}} \boldsymbol{\omega}^\top \Phi(\mathbf{I}; \mathbf{h}) = \max_{\mathbf{h}} \sum_{i=1}^{T} \omega_i^\top \phi(I_i; h_i) \qquad (1)$$

The model parameters $\boldsymbol{\omega}$ are simply the concatenation of the parameters for all temporal segments, i.e. $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_T)$, where w_i is the model parameter for temporal segment I_i. $\omega_i^\top \phi(I_i; h_i)$ can be interpreted as a score for the action of interest at spatio-temporal volume h_i.

4 Learning and Inference

In this section, we describe how to infer the action label given a video instance (Sec. 4.1) and how to learn the model parameters from the training set (Sec. 4.2). We perform binary classification in our experiments, i.e. $\mathcal{Y} = \{+1, -1\}$.

4.1 Inference

For each temporal segment I_i, the potential function $\omega_i^\top \phi(I_i; h_i)$ measures the compatibility between the action label $y \in \mathcal{Y}$ and the spatio-temporal volume $h_i \in \mathcal{R}_i$ in this temporal segment. The global scoring function $\boldsymbol{\omega}^\top \Phi(\mathbf{I}; \mathbf{h})$ measures the compatibility between the action label y and the whole video data. Given a test video $\mathbf{I}$ and the model parameters $\boldsymbol{\omega}$, the inference problem is the maximization of the scoring function $f_\omega(\mathbf{I})$ in Eq. (1) over all the possible latent variables $\mathbf{h} \in \mathcal{H}$.

Our latent structure is a temporal chain with the latent variables corresponding to the spatio-temporal volumes containing the person. Hence, the inference problem is essentially finding the maximum over all possible state sequences in a HMM, which is efficiently solved by the Viterbi algorithm in $O(T \times S^2)$ time.

4.2 Learning

Given a set of N positive samples $\{\mathbf{I}^i\}_{i=1}^N$ and M negative samples $\{\mathbf{I}^j\}_{j=N+1}^{N+M}$, we want to learn the model parameter $\boldsymbol{\omega}$ that tends to correctly predict the action label y and localize the person performing the action for a new test video $\mathbf{I}$.

We adopt the latent SVM framework [15, 5] for parameter learning, and consider the following optimization problem:

$$\mathcal{P}(\boldsymbol{\omega}^*) = \min_{\omega,\xi} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C_1 \sum_i \xi_i + C_2 \sum_j \xi_j \quad (2a)$$

$$s.t. \quad \boldsymbol{\omega}^\top \Phi(\mathbf{I}^i; \mathbf{h}^i) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (2b)$$

$$-\boldsymbol{\omega}^\top \Phi(\mathbf{I}^j; \mathbf{h}) \geq 1 - \xi_j, \quad \xi_j \geq 0 \quad (2c)$$

$$\forall i \in \{1, 2, \ldots, N\} \quad (2d)$$

$$\forall j \in \{N+1, \ldots, N+M\} \quad (2e)$$

where $\{\xi_i\}$ and $\{\xi_j\}$ are the slack variables for handling misclassification of difficult or noisy samples. The learning algorithm alternates between inferring $\mathbf{h}$ and optimizing $\boldsymbol{\omega}$. For the negative samples, we and select the most violated constraint over all possible vectors $\mathbf{h} = (h_1, h_2, \ldots, h_T)$ (Eq. (2c)). Similarly, we use the Viterbi algorithm to find the optimal value efficiently.

We use the non-convex bundle method [4] to solve Eq. (2). The algorithm iteratively builds an increasingly accurate piecewise quadratic approximation to the objective function based on bundle methods and cutting planes. Detailed explanations are omitted due to space constraints.

5 Experiments

We collected a dataset of real-world video footage to evaluate the performance of our model. This dataset comes from hundreds of hours of surveillance video data collected from long-term care facilities. Typical actions in this dataset include *walking*, *bending*, *standing* and *falling*. We selected the two most common actions, *falling* and *walking*, for evaluation. If a video contains a falling person, it is labeled as *fall*, otherwise *non-fall*. Similarly, if a video contains a walking person, it is labelled as *walk*, otherwise *non-walk*. We use 123 short clips containing 40 *fall* actions, 47 *walk* actions and 36 other actions. Each clip has 120 frames with frame size 320×240 pixels. We perform binary classification for *fall* versus *non-fall* and *walk* versus *non-walk*.

Our work on action localization and recognition is directly inspired by the potential application of fall detection and analysis in nursing home surveillance videos. Our clinician collaborators are studying the primary causes of real-life falls in high-risk environments, e.g. nursing homes, in order to design preventative strategies. Generally, *non-fall* video data are uninteresting for the purpose of fall analysis. Our model can be adopted to sift through these data to find potential instances of falls.

We divide each video clip into 3 temporal segments ($T = 3$), and each segment consists of 40 frames. We set the spatial parameter $S = 24$, which means there are 24 regions in one temporal segment; each region has size 120×120 pixels. We set $C_1 = C_2 = 10$ in Eq. (2) for all the experiments. We set the distance threshold $D = 170$ for *fall* actions and $D = 340$ for *walk* actions. The proposed model is compared with three baselines in the experiments:

1. **Holistic HOG3D**: The first baseline is a standard SVM classifier on a histogram of k-means quantized HOG3D descriptors [7] extracted by dense sampling from the entire video volume. We use a 3,000 word codebook in the experiments.
2. **Holistic Action Bank**: The second baseline is a standard SVM classifier on the Action Bank descriptor computed for the entire video volume.
3. **Latent localization**: The third baseline is in the same framework of our proposed model. The only difference is that we set the temporal parameter $T = 1$ – a single latent variable h to represents the discriminative region containing the person performing the action.

5.1 Experimental Results

We summarize the comparison of our model and the baselines in Table 1. We can see that our model significantly outperforms the baseline methods. The first two baselines use a standard SVM framework without introducing latent variables. The third baseline is in the framework of our proposed model but without considering temporal variation.

Action	Method	Accuracy
fall	holistic HOG3D	65.00%
	holistic Action Bank	67.50%
	latent localization	82.50%
	our model	**87.50%**
walk	holistic HOG3D	63.04%
	holistic Action Bank	71.74%
	latent localization	82.61%
	our model	**86.96%**

Table 1. Comparison of action classification accuracies of different methods on the Nursing Home Dataset. We tested these methods with two action labels *fall* and *walk*.

We visualize the action localization results with discriminative regions (latent variables $\mathbf{h}$) for *fall* and *walk* actions in Fig. 5.

We can see the effectiveness of introducing latent localization into our framework. Evidence for this is provided by the performance improvement over the baseline methods. This indicates the latent localization of the person performing the action is helpful for learning discriminative model parameters. In addition, temporal links based on the HMM-like structure improve the performance for recognizing complex actions. The improvement is larger for *fall* actions, which likely contain greater variation over time.

Typical misclassified examples for *fall* actions include *bending* and *sitting* actions, e.g. people who quickly bend down to pick up trash are predicted as *fall* in the test videos. These actions share similarities with *fall* actions. Misclassified examples for *walk* actions include people walking in directions that lack sufficient examples in the training data.

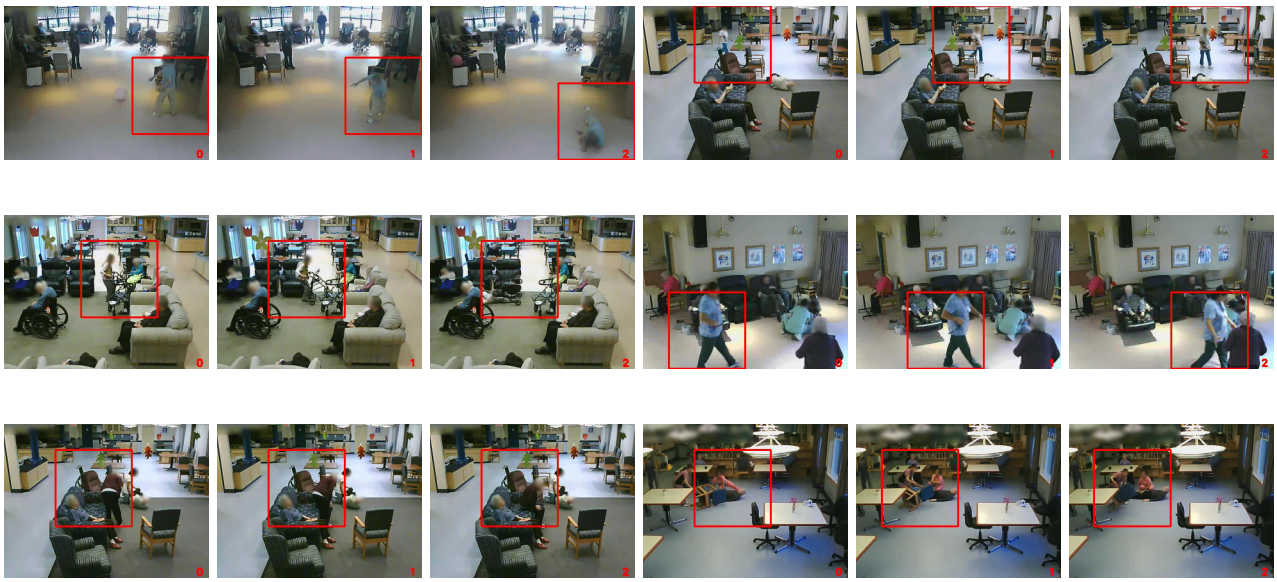

Figure 3. (Best viewed in color) Visualization of action localization results for *fall* and *walk* actions. The bounding boxes are the discriminative regions (or equivalently latent variables **h**). One frame is retrieved from each temporal segment. The first two rows show correct examples, which are correctly predicted by our models but not by the baselines. The third row shows incorrect examples. Left shows *fall* label on a person who sits down. Right shows video with label *walk* on a person who moves a chair. Faces blurred for privacy.

6 Conclusion

We presented a discriminative latent spatio-temporal model for action localization and recognition. The proposed model does not require human detection to initialize the inference process, and uses a rich Action Bank feature representation. We develop a latent variable framework, which treats spatio-temporal locations of the person as latent variables. Our experimental results demonstrate that our proposed model significantly outperforms baseline methods, and shows promise for automatic detection of fall events in real-world care facility videos.

Acknowledgement

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Canadian Institutes of Health Research (CIHR; grant numbers AMG-100487 and TIR-103945).

References

[1] M. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *ECCV*, 2012.

[2] H. Bilen, V.P. Namboodiri, and L.J. Van Gool. Object and Action Classification with Latent Variables. In *Proceedings of The British Machine Vision Conference*, 2011.

[3] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proc. of European Conference of Computer Vision*, 2012.

[4] T.-M.-T. Do and T. Artières. Large margin training for hidden markov models with partially observed states. In *ICML*, 2009.

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

[6] Z. F. Huang, W. Yang, Y. Wang, and G. Mori. Latent boosting for action recognition. In *22nd British Machine Vision Conference (BMVC)*, 2011.

[7] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, 2008.

[8] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *International Conference on Computer Vision (ICCV)*, 2011.

[9] S. N. Robinovitch, F. Feldman, Y. Yang, R. Schonnop, P. M. Leung, T. Sarraf, J. Sims-Gould, and M. Loughin. Video capture of the circumstances of falls in elderly people residing in long-term care: an observational study. *The Lancet*, 381:47–54, 2013.

[10] S. Sadanand and J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[11] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

[12] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *CVIU*, 115:224–241, 2011.

[13] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 1992.

[14] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *CVPR*, 2010.

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