Recurrent Attention Models for Depth-Based Person Identification

Albert Haque, Alexandre Alahi, Li Fei-Fei
Computer Science Department, Stanford University
{ahaque,alahi,feifeili}@cs.stanford.edu

Abstract

We present an attention-based model that reasons on human body shape and motion dynamics to identify individuals in the absence of RGB information, hence in the dark. Our approach leverages unique 4D spatio-temporal signatures to address the identification problem across days. Formulated as a reinforcement learning task, our model is based on a combination of convolutional and recurrent neural networks with the goal of identifying small, discriminative regions indicative of human identity. We demonstrate that our model produces state-of-the-art results on several published datasets given only depth images. We further study the robustness of our model towards viewpoint, appearance, and volumetric changes. Finally, we share insights gleaned from interpretable 2D, 3D, and 4D visualizations of our model's spatio-temporal attention.

1. Introduction

A quick, partial view of a person is often sufficient for a human to recognize an individual. This remarkable ability has proven to be an elusive task for modern computer vision systems. Nevertheless, it represents a valuable task for security authentication, human tracking, public safety, and role-based activity understanding [34, 30, 2].

Given an input image, person identification aims to assign identification labels to individuals present in the image. Despite the best efforts from previous work [79, 80, 40], this problem remains largely unsolved. Without accurate spatial or temporal constraints, visual features alone are often intrinsically weak for matching people across time due to intra-class differences. Additional variances due to illumination, viewpoint, and pose further exacerbate the problem.

Research findings from physiology and psychology have shown that gait is unique to each individual [57, 56, 17]. Building on this observation, we aim to learn body shape and motion signatures unique to each person (see Figure 1). Inspired by the recent success of the depth modality [4, 77], our goal is to output an identification label from a depth image or video.
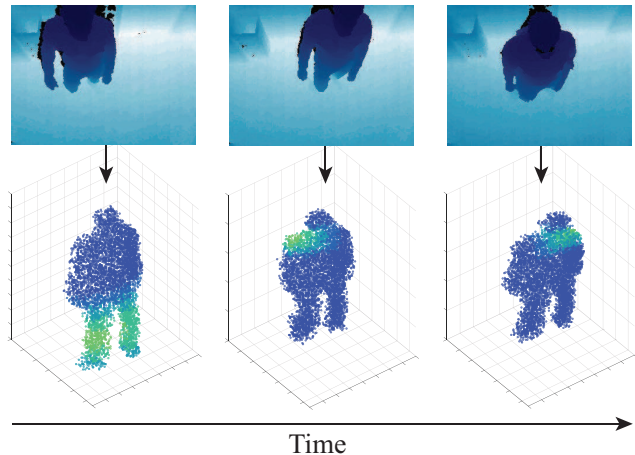

Figure 1: Gait has been shown to be unique to each person. We propose a 4D recurrent attention model to learn spatio-temporal signatures and identify people from depth images.

The primary challenge towards this goal is designing a model that is not only rich enough to reason about motion and body shape but also robust to intra-class variability. The second challenge is that person identification inherently comprises of a large number of classes with few training examples per class (in some cases a single training example). Existing datasets [52, 5, 54] often collect front-facing views with constant appearances (*i.e.* similar sets of clothing). While this makes the identification problem more tractable, we are interested in relaxing these assumptions to solve a more general identification task which is applicable to a broader audience.

Our core insight is that we can leverage raw depth video despite the scarcity of training inputs, to address the aforementioned challenges by formulating the task as a reinforcement learning problem. Our approach involves pruning the high dimensional input space and focuses on small, discriminative regions while being free of visual and temporal assumptions. Concretely, our contributions are:

(i) We develop a recurrent attention model that identifies humans based on depth videos. Our model leverages a 4D

input and is robust to appearance and volumetric changes. By combining a sparsification technique with a reinforcement learning objective, our recurrent attention model attends to small spatio-temporal regions with high fidelity while avoiding areas with little information (see Section 3).

(ii) We re-examine the person identification task and build a challenging dataset which taxes existing methods (see Section 4). We push the limits of our model by varying the viewing angle and testing on diverse training examples of people carrying objects (*e.g.* coffee or laptops) or wearings hats and backpacks.

In Section 4, we show that our model achieves state-of-the-art results on several existing datasets. Furthermore, we take advantage of our recurrent attention model and create interpretable 2D, 3D, and 4D visualizations of hard attention [71]. Our findings shed new insights on volumetric and motion-based differences between individuals. To aid in future research, we make all code, data, and annotations publicly available upon publication.

2. Related Work

RGB-Based Methods. The primary challenge associated with identification is intra-class variance. These include changes in appearance due to illumination, point of view, pose, and occlusion. There have been many attempts to solve this problem by improving the feature representations [24, 68, 22, 78, 79, 37, 80] and by exploring new similarity metrics [40, 49, 60]. Silhouette-based approaches ignore color altogether and use anthropometric or geodesic distances between body parts [34, 44, 64].

Depth-Based Methods. Following suit from silhouette-based approaches, several depth-based studies have applied anthropometric and soft biometrics to the 3D human skeleton [51, 3, 55, 4, 20]. Harnessing the full power of depth cameras, several papers investigated 3D point clouds for person identification [77, 30]. Although these approaches are successful, they rely on hand-crafted features (*e.g.* arm length, torso width) or low-level RGB features (*e.g.* SURF [7], SIFT [42]).

Spatio-Temporal Representations. Methods described thus far have largely ignored spatio-temporal information. Originally proposed in [26], the gait energy image and ifigts variants [16, 6, 29, 66], embed temporal information onto a two-dimensional image by averaging the silhouette across all frames of a video. Test time predictions are obtained from a k-nearest neighbor lookup.

More recently, the gait energy image has been extended into 3D by using depth sensors [28, 63]. Spatial volumes and higher-dimensional tensors have been proposed for activity and action recognition [58, 67, 75, 8, 32, 39], medical image analysis [62], robotics [47, 48], and human motion analysis [38] but have not been thoroughly explored in the person identification domain.

Deep Learning for Identification. A small number of studies have explored the applicability of deep neural networks to person identification. In [73], Yi et al. proposed a siamese convolutional neural network for similarity metric learning. In [41], Li et al. proposed a similar approach by using filter pairs to model photometric and geometric transforms. Following these works, Ding et al. [19] formulated the input as a triplet containing both correct and incorrect reference images. In [1], Ahmed et al. introduced cross-input neighborhood differences.

Our work has several key differences with the aforementioned works: First, we focus on the depth modality and do not use any RGB information. Second, the methods above [73, 41, 19, 1] ingest several images as input and compute similarity between these inputs. They formulate the identification problem as an image-similarity task using images captured from non-overlapping camera views. Our model uses a single image[1] as input and does not rely on metric learning.

Attention Models. Interpretability of deep learning models is becoming increasingly important within the machine learning and computer vision communities. By measuring the sensitivity of output variables to variances in the input, attention models applied to image classification [76, 25, 70], image captioning [21, 71, 14], object detection [11], and tracking [18] have demystified many aspects of convolutional and recurrent networks. These methods exploit the spatial structure of the input to understand intermediate network representations. Sequential data, on the other hand, requires temporal attention models to understand the order dependence of the input data. Recent papers in speech recognition [23], video captioning [72], and natural language processing [35, 43, 13] explore the concept of attention in the temporal domain.

Many deep learning models impose constraints on the input. Due to the high dimensionality of images (*i.e.* high pixel count), preprocessing often includes resizing and/or cropping the original input image [36]. Videos are often truncated to a fixed length for training. Due to computational limitations, this loss of information is necessary to constrain runtimes. In the next section, we describe our model and how we balance this trade-off by employing visual "glimpses" [50] which process small 4D regions with high fidelity and grow to larger regions with lower detail.

3. Our Model

The goal of our model is to identify humans from depth images or video. Our model (Figure 2) computes hard attention regions [71] which are used to predict an identification label. In this section, we describe our 4D input representation followed by a discussion of our attention model.

---

[1]The input to our model is one image for frame-wise identification or one sequence for video-level (i.e. temporal or voting) identification.
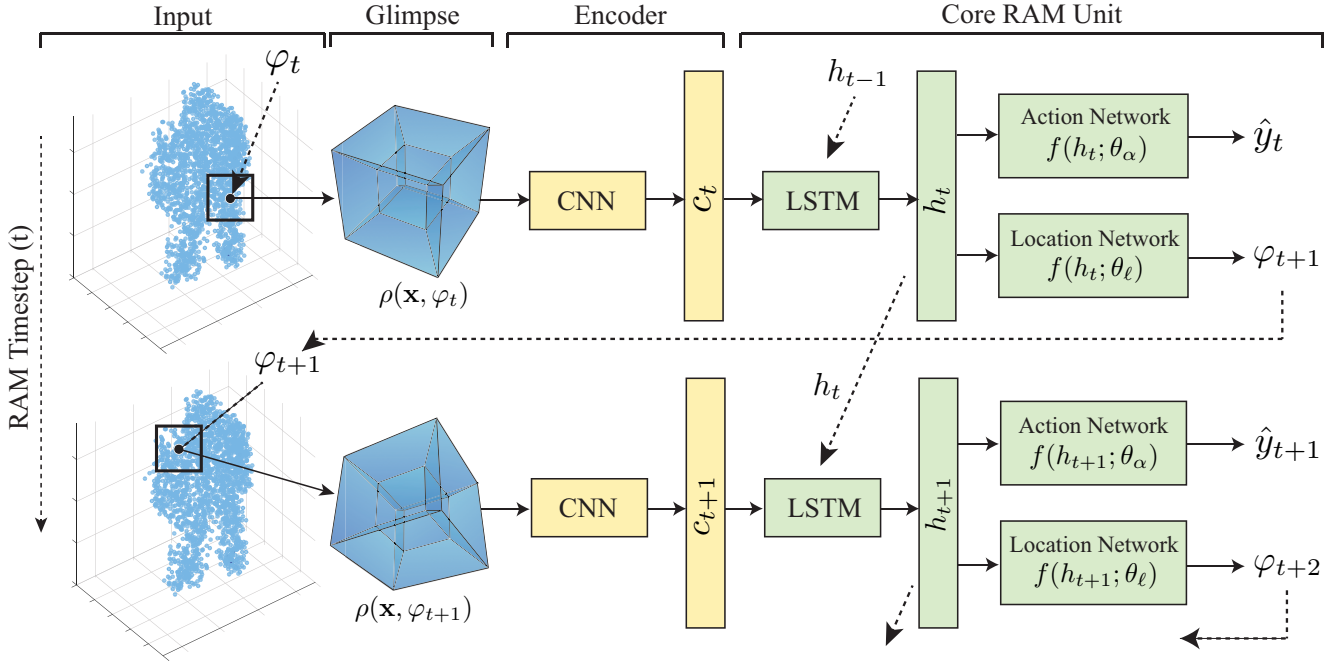
Figure 2: Our full model. Dashed arrows indicate information exchange across time steps. Solid arrows indicate information exchange within a time step. Two time steps are shown with a series of events occurring from left to right. Note: RAM timestep t refers to the "iteration" of our model and does not refer to the input video timestamp τ. All other variables are defined in Section 3.2.

3.1. Input Representation

Projections from higher dimensional spaces onto lower spaces result in information loss. This serves as our motivation for using 4D data: we want to preserve as much information as possible and let our model decide the relevant regions. Four-dimensional data consists of a 3D point cloud (*e.g.*, x, y, and z cooridnate) and time τ. For simplicitly, Figure 2 shows the input as 3D point clouds which are constructed from depth images.

Each training example $(\mathbf{x}, \mathbf{y})$ consists of a variable sized 4D tensor $\mathbf{x}$ and corresponding label $\mathbf{y}$. The tensor is variable due to variable video lengths. Let f denote the number of frames in video i and let x, y and z denote the width, height, and depth dimensions of our tensor.[2]

$$\mathbf{x} \in \mathbb{R}^{f \times x \times y \times z} \quad \text{and} \quad \mathbf{y} \in [1, ..., C] \qquad (1)$$

where C is the number of classes. For an average video containing 500 frames, flattening $\mathbf{x}$ leads to a feature vector of 2.5×10^9 elements. For comparison, a 227×227 RGB image (typical for a convolutional network), results in 1.2×10^6 elements. This means that our model must operate on an input space three orders of magnitude larger than common convolutional networks. Consequently, our model must be designed to intelligently navigate this high dimensional space.

---

[2]We use a tensor of size $250 \times 100 \times 200$.

3.2. Recurrent Attention Model

Given this high-dimensional depth representation, we want our model to focus on smaller, discriminative regions in the input space. Minh et al. [50] recently proposed the recurrent attention model (RAM) for image classification and reinforcement learning problems. While they show promising results, they enjoyed several advantages. First, training data is plentiful. Image classification has been well-studied and several large benchmarks exist. Dynamic environments such as a control-based video game can generate data on-the-fly as the game is played. Second, the input dimensionality of these problems is relatively small: MNIST is 28×28 while the control game is 24×24 [50].

Person identification, on the other hand, does not enjoy these advantages. Instead, we are tasked with limited, high-dimensional training data. Figure 2 shows an overview of our proposed model. It consists of a glimpse layer which down-samples the input, an encoding stage which acts as an additional dimensionality reduction tool, and a core RAM network responsible for spatio-temporal learning.

Glimpse Layer. The goal of the glimpse layer is two-fold: (i) it must avoid (or greatly limit) information loss and (ii) it must refrain from processing large inputs. At a given time step t, our model does not have full access to the input $\mathbf{x}$ but instead extracts a partial observation or "glimpse" denoted by $\rho(\mathbf{x}, \varphi_t)$. A glimpse encodes the region around

φ_t with high resolution but uses a progressively lower resolution for points further from φ_t. Adopting a multi-scale strategy has been shown to be an effective de-noising technique [81]. Additionally, this results in a tensor with much lower dimensionality than the original input $\mathbf{x}$. By focusing on specific regions, we can reduce the required computation by our model, reducing the loss of spatio-temporal detail, and reduce the effect of noise.

As shown in Figure 2, a glimpse is comprised of G hypercube patches. The first patch has a side length of g_s and maintains full resolution of the input centered at φ. The second patch has a side length of $2g_s$ and is sampled at $1/2$ resolution. Patches grow in size with progressively lower resolution. Specifically, the kth patch has a side length of kg_s and is sampled at $1/k$ of the original input resolution. The final glimpse is a concatenation of these hypercube patches.

Encoder. The glimpse still contains a large number of features (on the order of 1×10^6). We must further compress the glimpse before it becomes a feasible solution for our data-limited person identification task. To accomplish this, we use an encoding layer to further reduce the feature space. In our model, this is done with a 4D convolutional autoencoder [45, 33]. The encoder layer is trained offline and separately from the RAM. During RAM training and test time, encoded features are denoted as c_t.

Core RAM Unit. As mentioned previously, the number of features associated with a 4D input is on the order of 1×10^9. Conventional deep learning methods cannot feasibly explore and learn from the full input space. Motivated by this, we use a recurrent attention model. Our goals of the RAM are two-fold: First, model interpretability is an overarching theme of this work. Given image-based input, an attention-based model allows us to visually understand human shape and body dynamics. Second, a RAM provides us with computational advantages by pruning the input space by focusing on rich, discriminative regions.

As shown in Figure 2 our model is a recurrent network: it consists of a long short-term memory (LSTM) unit [27] and two sub-networks. Parameterized by θ_r, our LSTM receives encoded features c_t and the previous hidden layer h_{t-1} at each time step t and outputs a hidden state h_t.

Sub-Networks. Before the next iteration of our RAM, our model must take two actions: (i) it decides the next glimpse location and (ii) it outputs a predicted identification label for the current time step. We compute these by using two sub-networks: the location and action network, respectively.

The location network stochastically selects the next glimpse location using the distribution parametrized by $f(h_t; \theta_\ell)$ (where θ_ℓ refers to the location network's parameters). Similar to [50], the location network outputs the mean of the location policy (defined by a 4-component Gaussian) at time t and is defined by: $f(h_t; \theta_\ell) = tanh(Linear(h_t))$

where $Linear(\bullet)$ is a linear transformation.

The action network (parameterized by θ_α), outputs a predicted class label $\hat{y}$ given the current LSTM hidden state, h_t. Parameterized by $f(h_t; \theta_\alpha)$, the action network consists of a linear and softmax layer defined by $f(h_t; \theta_\alpha) = \exp(Linear(h_t))/Z$ where Z is a normalizing factor. The predicted class label $\hat{y}_t$ is then selected from the softmax output.

3.3. Training and Optimization

Formulation. Depth video is inherently a large feature space. To avoid exploring the entire input space, we pose the training task as a reinforcement learning problem. After our model decides the label $\hat{y}$ and next glimpse location φ, our model receives a reward R where $R = 1$ if $\hat{y}_t = \mathbf{y}$ at time T, where T is a threshold for the maximum number of time steps; otherwise $R = 0$. Let $\Theta = \{\theta_r, \theta_\ell, \theta_\alpha\}$ denote all parameters of the RAM.

Let $s_{1:t} = \mathbf{x}, \varphi_1, \hat{y}_1, ..., \mathbf{x}, \varphi_t, \hat{y}_t$ denote the historical sequence of all input-action pairs (*i.e.* input tensor, predicted label, and next glimpse). We call this a *glimpse path*. A glimpse path shows where our model "looks at" over time[3]. Our model must learn a stochastic policy $\pi(\varphi_t, \hat{y}_t | s_{1:t}; \Theta)$ which maps the glimpse path $s_{1:t}$ to a distribution over actions for the current time step. The policy π is defined by our core RAM unit and the history s_t is embedded in the LSTM's hidden state h_t.

Optimization. The policy of our model induces a distribution over possible glimpse paths. Our goal is to maximize the reward function over $s_{1:N}$:

$$J(\Theta) = \mathbb{E}_{p(s_{1:t}; \Theta)}[R] \quad (2)$$

where $p(s_{1:T}; \Theta)$ depends on the policy π. However, computing the expectation introduces unknown environment parameters which makes the problem intractable. Formulating the task as a partially-observable Markov decision process allows us to compute a sample approximation to the gradient, known as the REINFORCE rule [69]:

$$\nabla_\Theta J(\Theta) = \sum_{t=1}^{T} \mathbb{E}_{p(s_{1:T}; \Theta)}\left(\nabla_\Theta \log \pi(\mathbf{y} | s_{1:t}; \Theta) R\right) \quad (3)$$

$$\approx \frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} \nabla_\Theta \log \pi(\mathbf{y} | s_{1:t}^{(i)}; \Theta) R^{(i)} \quad (4)$$

where $s_{1:t}^{(i)}$ denotes the glimpse path, $R^{(i)}$ denotes the reward, and $\mathbf{y}^{(i)}$ denotes the correct label for the ith training example. Additionally, $\nabla_\theta \log \pi(u_t^{(i)} | s_{1:t}^{(i)}; \theta) R^{(i)}$ is the gradient of the LSTM. Consistent with [50], we train the action network with the cross entropy loss function and train the

---

[3]For 4D input, time refers to the iteration of the RAM and not the input video's frame order.

	BIWI	IAS-A/B	PAVIS	DPI-T
# Unique Subjects	50 (28)	11 (11)	**79 (79)**	12 (12)
# Total Videos	50 (56)	11 (11)	79 (79)	**300 (355)**
# Appearances/class	1 (2)	1 (2)	1 (2)	**5 (5)**
# 2D inputs/class	479 (551)	**701 (739)**	5 (5)	336 (398)
# 3D inputs/class	479 (551)	**701 (739)**	5 (5)	336 (398)
# 4D inputs/class	1 (2)	1 (1)	1 (1)	**25 (30)**

Table 1: Comparison of datasets. DPI-T is our newly collected dataset. We list the number of subjects, images, and videos for both the training and test sets. The test set is shown in parenthesis. Appearance is defined by a person wearing unique clothing or distinct visual appearance.

location network with REINFORCE. This formulation allows our model to focus on salient 3D regions in both space and time.

Advantages. A major benefit of this formulation is that limited training data is no longer an issue. Our model is trained on glimpses (*i.e.* subsets of the input) and not the entire video sequence. Therefore, the effective number of training examples made available to our model is on the order of 1×10^6 to 1×10^9 per video (*i.e.* number of possible glimpses). Despite having a single video as input, our model almost never sees the same training example twice. Our model is still limited by the number of training data but our formulation makes it less of a concern.

4. Experiments

First, we describe our datasets and evaluation metrics. This is followed by a discussion of experimental, hyperparameter, and design selections. We then present results for the single-shot (single image) and multi-shot (multiframe) person identification task. We then show 2D, 3D, and 4D visualizations followed by concluding remarks on our model's limitations.

4.1. Datasets

Our goal is to identify humans based on their 3D shape and body dynamics captured by a depth camera. The majority of human-based RGB-D datasets are catered to human activity analysis and action recognition [12, 10, 31]. Since they generally consist of many gestures performed by few subjects, these datasets are not suited for the identification problem. We hence use existing depth-based identification datasets and collected a new one to further test our model.

We evaluate our model on several existing depth-based identification datasets: BIWI [52], IIT PAVIS [5], and IAS-Lab [54]. These datasets contain 50, 79, and 11 humans, respectively. For BIWI, we use the full training set and the *Walking* test set. For PAVIS, we use *Walking1* and *Walking2* as the training and test set, respectively. For IAS-Lab, we use the full training set and both test splits.

Existing datasets impose constraints to simplify the identification problem (*e.g.*, few sets of clothing per person,
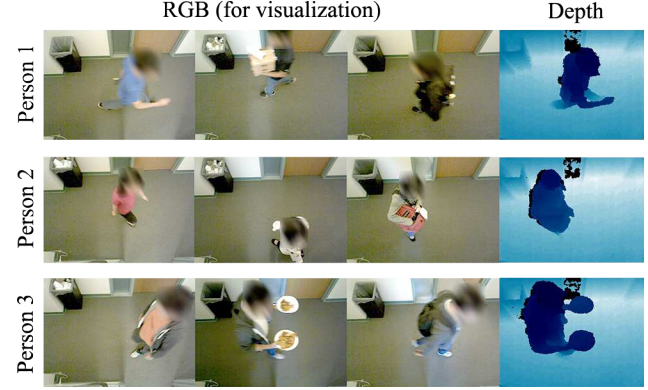

Figure 3: Sample images from our Depth-Based Person Identification from Top (DPI-T) dataset. Each row denotes a different person. The three left columns show RGB images for convenience. Our model only uses depth images, as depicted in the right column.

front-facing views, or slow walking speed). We collected a new dataset: Depth-Based Person Identification from Top (DPI-T), which is different from previous datasets.

We provide more observations per individual. On average, individuals appear in a total of 25 videos across several days. This naturally results in individuals wearing different sets of clothing – 5 different sets of clothing on average. Figure 3 shows three individuals from our dataset wearing different sets of clothing. Additionally, people in our dataset walk at variable speeds depending on the time of day or week.

Challenging top-view angles. In real-world applications such as smart spaces and public environments (*e.g.*, hospitals, retail stores), cameras are often attached to the ceiling pointed down, as opposed to clean, frontal or side view images available in existing datasets. This introduces self-occlusion challenges and often leads to undetected faces and incomplete 3D point cloud reconstructions.

People are holding objects. Existing datasets collect data from the simple case of walking in a controlled environment. In our dataset, people are "in the wild," often holding objects such as coffee, laptops, or food. Additionally, since our dataset is collected across a long period of time, people often wear hats, bags, or carry umbrellas (see Figure 3). A table showing the characteristics of existing datasets and our new dataset is shown in Table 1.

4.2. Evaluation Metrics

Person identification can be solved in a "single-shot" manner using one image to produce a label or a "multi-shot" method which leverages multiple frames, temporal features, or multi-frame voting schemes. Below, we provide evaluation results for both single-shot and multi-shot approaches.

#	Modality	Methods	Top-1 Recognition Rate (%)					Normalized Area Under the Curve (nAUC)				
			BIWI	IAS-A	IAS-B	PAVIS	DPI-T	BIWI	IAS-A	IAS-B	PAVIS	DPI-T
1	Depth	Random	2.0	9.1	8.1	1.3	8.3	51.0	54.5	54.5	50.6	54.2
2	Depth	Human Performance	6.7	21.2	15.1	1.7	19.2	—	—	—	—	—
3	Depth	Skeleton (NN) [5]	—	—	—	15.0	—	—	—	—	91.8	—
4	Depth	Skeleton (NN) [52]	21.1	22.5	55.5	28.6	—	81.7	72.8	86.3	89.9	—
5	Depth	Skeleton (SVM) [53]	13.8	—	—	35.7	—	86.6	—	—	92.8	—
6	Depth	3D CNN	27.7	44.2	56.2	27.5	23.7	88.2	86.1	86.0	89.2	75.6
7	Depth	2D RAM	24.7	46.9	61.0	30.5	33.8	87.4	87.7	86.8	90.1	82.5
8	Depth	3D RAM	**30.1**	**48.3**	**63.7**	**41.3**	**47.5**	**88.7**	**88.5**	**87.7**	**93.7**	**88.3**
9*	RGB	Face Detection [52]	36.7*	—	—	—	—	87.6*	—	—	—	—
10*	RGB	PTZ Max-Var [61]	—	—	—	73.1*	—	—	—	—	98.7*	—
11*	RGB-D	Face+Skeleton [53]	43.9*	—	—	—	—	90.2*	—	—	—	—
12*	RGB-D	PCM+Skeleton [52]	27.4*	25.6*	63.3*	—	—	87.4*	75.5*	86.3*	—	—

Table 2: Single-shot identification performance. Methods shown above use only spatial information. A summary of each method can be found in Section 4.4. Both metrics were computed on the test set. Larger values are better. Dashes indicate that no published information is available. (*) Although not a fair comparison, for sake of completeness, we list RGB and RGB-D methods.

(a) BIWI (Single-Shot) (b) IAS-Lab-A (Single-Shot) (c) IAS-Lab-B (Single-Shot)
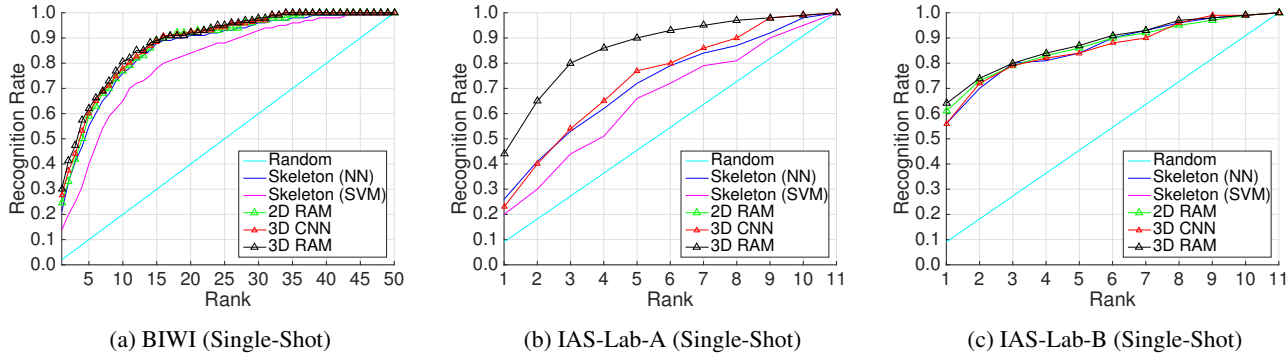
Figure 4: (a-c) Cumulative matching curves for test set performance on various datasets and models. Dataset details can be found in Section 4.1. Model details can be found in Section 4.4. The y axis denotes recognition rate. For the x axis, rank-k is the recognition rate if the ground truth label is within the model's top-k predictions.

Specific metrics include the top-1 recognition rate, cumulative matching curve (CMC), normalized area under the curve (nAUC) metrics. Top-k recognition rate indicates the fraction of test examples that contained the ground truth label within the top-k predictions. Generalizing the top-k metric to higher ranks (up to the number of people in the dataset), produces the cumulative matching curve. Integrating the area under the CMC curve and normalizing for the number of ranks produces the nAUC.

4.3. Experimental Settings

Tensors were fixed to a size of $250 \times 100 \times 200$ and converted to integer indices corresponding to the $x, y,$ and z real world coordinates. The x and y units represent real world centimeters while the z units represents 10 millimeters. Glimpse locations are encoded as $\varphi = (x, y, z, \tau)$ where x, y, z are real values while τ is integer valued. The first glimpse patch has a side length of 8 tensor units and we use 5 glimpse patches. For 3D and 4D inputs, we augment the data by applying Gaussian noise, with mean of 0 cm and 5 cm variance, to each point in the point cloud. Images

and tensors are shifted between 0 and ± 5 cm in all directions about the origin and randomly scaled between $0.8\times$ and $1.2\times$. We train our model from scratch using stochastic gradient descent with mini-batches of size 20, a learning rate of 1×10^{-4}, momentum of 0.9, and weight decay of 5×10^{-4}. The CNN was pretrained on augmented training examples before RAM training. All learning layers employ dropout [65] with 0.5 probability.

4.4. Baselines

Single-Shot Identification. We compare our recurrent attention model to several depth-based methods. Table 2 shows various methods and results for the single-shot identification task: (1) We computed performance using a uniformly random guessing strategy. (2) Four humans manually performed the identification task. Each human was shown a single test input and was given full access to the training data. (3-5) Distances between skeleton joints are used as hand-crafted features [5, 52, 53]. (6) A three-dimensional CNN operates on 3D point clouds. (7) A two-dimensional RAM operates on depth images. (8) A three-

#	Modality	Methods	Top-1 Recognition Rate (%)					Normalized Area Under the Curve (nAUC)				
			BIWI	IAS-A	IAS-B	PAVIS	DPI-T	BIWI	IAS-A	IAS-B	PAVIS	DPI-T
1	Depth	Random	2.0	9.1	8.1	1.3	8.3	51.0	54.5	54.5	50.6	54.2
2	Depth	Human Performance	6.7	21.2	15.1	1.7	19.2	—	—	—	—	—
3	Depth	Energy Image [16]	21.4	25.6	15.9	29.1	18.5	73.2	72.1	66.0	81.2	75.8
4	Depth	Energy Volume [63]	25.7	20.4	13.7	18.9	14.2	83.2	66.2	64.8	68.3	65.5
5	Depth	Skeleton (NN) [53]	39.3	—	—	—	—	—	—	—	—	—
6	Depth	Skeleton (SVM) [53]	17.9	—	—	—	—	—	—	—	—	—
7	Depth	Skeleton (LSTM)	15.8	20.0	19.1	14.5	—	65.8	65.9	68.4	64.0	—
8	Depth	3D CNN+Avg Pooling [9]	27.8	33.4	39.1	27.5	28.4	84.0	81.4	82.8	80.6	82.5
9	Depth	3D LSTM	27.0	31.0	33.8	20.3	23.9	83.3	77.6	78.0	77.1	77.9
10	Depth	4D RAM	**45.3**	**53.5**	**64.4**	**43.0**	**55.6**	**91.2**	**91.4**	**89.0**	**93.4**	**91.6**
11*	RGB	Face Detection [52]	57.1*	—	—	—	—	—	—	—	—	—
12*	RGB-D	Face+Skeleton [53]	67.9*	—	—	—	—	—	—	—	—	—
13*	RGB-D	MCL+Skeleton [59]	—	—	—	89.0*	—	—	—	—	98.9*	—
14*	RGB-D	PCM+Skeleton [52]	42.9*	27.3*	81.8*	—	—	—	—	—	—	—

Table 3: Multi-shot identification performance. Methods shown above use multiple test images or use temporal information. A summary of each method can be found in Section 4.4. Both metrics were computed on the test set. Larger values are better. Dashes indicate that no published information is available. (*) Although not a fair comparison, for sake of completeness, we list RGB and RGB-D methods.
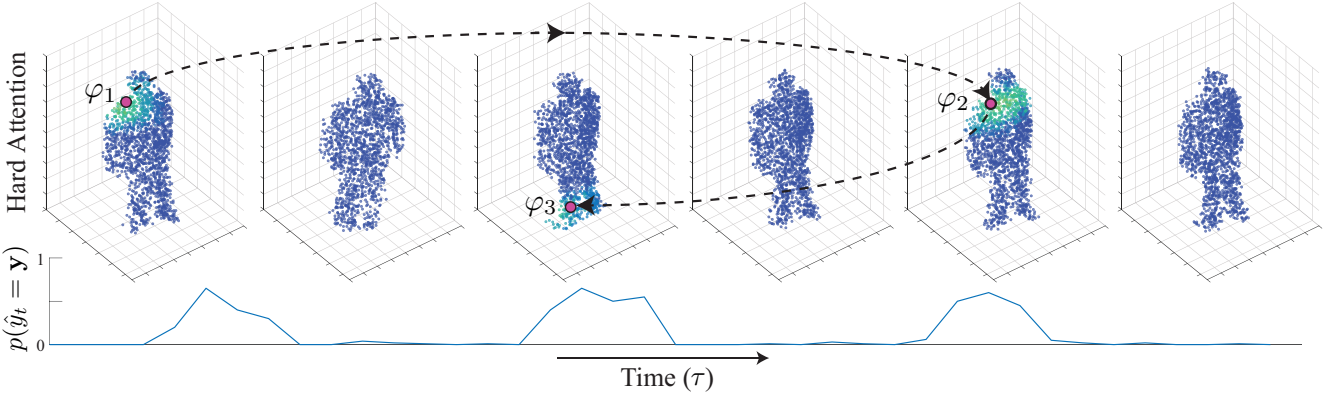

Figure 5: Hard 4D attention regions. Bright colored regions indicate areas closer to the glimpse center. Each point cloud is shown above in three dimensions (x, y, z) while the point clouds are arranged in video-order starting from the left. Arrows indicate jumps in time τ. Although sparse point clouds are shown above, our model operates on the raw, dense point clouds.

dimensional RAM operates on 3D point clouds. Although the focus of our paper is depth-based person identification, for completeness, we include related RGB and RGB-D methods to provide a more holistic view of the field. (9) A face descriptor is used [52]. (10) A point-tilt-zoom camera selectively zooms in on different parts of the image [61]. (11) A facial descriptor is concatenated with distances between skeleton joints [53]. (12) Similarity scores are computed on 3D point clouds and distances between skeleton joints [52].

Multi-Shot Identification. Table 3 list several multi-shot methods. (1-2) We use random and human performance as baselines. (3-4) We evaluate the gait energy image [26] and volume [63]. (5-6) These methods use hand-crafted skeleton features with an inter-frame voting system. (7) Pairwise skeleton joint distances (same as 5-6) are fed into a LSTM. (8) A 3D CNN with average pooling [9] over time [74] . (9) A 3D LSTM operates on 3D point clouds. (10) Our final RAM model. (11-12) Face descriptors are

used with a voting system. (13) A Multiple Component Dissimilarity (MCD) metric is computed on a pair of images. (14) RGB-D point cloud matching plus hand-crafted features are used for identification.

4.5. Single-Shot Identification Performance

Learned encoding improves performance. To better understand the source of our performance, we reduced the input dimensionality of our RAM and evaluated a 2D and 3D variant. The 2D and 3D models were evaluated on the single-shot task. As the dimensionality of the input increases from 2D to 3D, the performance of our RAM monotonically increases (see Figure 4). Contrast this with gait energy in Table 2. Gait energy undergoes a similar transformation from 2D to 3D (i.e. image to volume), but exhibits lower performance in the higher-dimensional case. This indicates that our learned encoder is able to preserve pertinent information from higher dimensional inputs whereas the gait energy volume fails without such encoding.

RAM outperforms deep learning baselines. As further validation of our model's performance, we evaluated a 3D convolutional neural network [33]. The input to both the 3D CNN and 3D RAM are 3D point clouds. As shown in Table 2, our 3D RAM model outperforms the 3D CNN. This confirms our hypothesis that our RAM is able to leverage glimpses to artificially increase the number of training examples and improve performance. The 3D CNN does not perform such data augmentation and instead operates on the entire point cloud.

4.6. Multi-Shot Identification Performance

Our final model (4D RAM) outperforms the human baseline and existing depth-based approaches. Both Munaro *et al*. [52] and Barbosa *et al*. [5] used distances between skeleton joints as features. We list the performance of these hand-crafted features in Table 3. Results show that these features are unable to infer the complex latent variables. Our 4D RAM model also outperforms an RGB-D method (13) in Table 3. Proposed in [52], method (13) computes a standardized 3D point cloud representation with the above skeleton-distance features. Although (13) leverages RGB information, it analyzes the entire point cloud which may include extraneous noise. Our model avoids noisy areas by selecting glimpses which contain useful information.

4.7. Hard Attention Regions

There is one key difference between our 3D and 4D RAM. In the 3D case, our model must "pay attention" to regions for each frame τ. However, in the 4D case, our model does not have this requirement since τ is a free parameter. Our model has full discretion on which frames to "pay attention to" and can move both forward and backward in time as needed. We analyze this in Figure 5. Over the course of the video, $p(\hat{y}_t = \mathbf{y})$ varies. Not only can our model change the glimpse's spatial location in each frame, it can also change the magnitude. Although our model has no explicit notion of attention magnitude, it can indirectly mimic the concept. To reduce the magnitude of attention given to frame k, our model moves the glimpse center to a frame further away from k. Although the overal "magnitude" of attention remains constant for each glimpse, the amount of attention given to k has been reduced.

As shown in Figure 5, our model begins at φ_1, "looks at" the person's shoulder, jumps to a different frame, and continues "staring at" the shoulder. One interpretation of this is that our model has learned to identify periodic cycles. Interestingly, it has been shown in the biological literature that males exhibit strong rotational displacement at the shoulders while walking [46]. Our model's attention corroborates this claim. The model then jumps backward in time and attends to the feet at φ_3. This indicates that leg motions (*i.e*. gait) potentially provide traces of identity. It
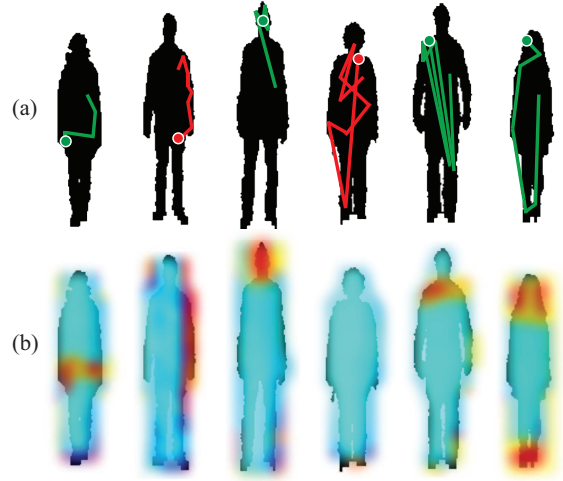

Figure 6: Two-dimensional projections of our model's 4D attention. (a) Glimpse paths. Green and red lines indicate correct and incorrect class label predictions, respectively. Circles denote the final glimpse location which led to the prediction. (b) Glimpse heatmap. Red regions denote areas on the human body frequently visited by our model. Heatmaps were smoothed with a Gaussian filter.

is quite possible that this particular glimpse path was taken since our learned policy simply never explored other paths, but our model was trained over many epochs with different initial glimpse locations to reduce this possibility.

We then project the 4D attention onto a 2D image. Figure 6a shows glimpse paths taken by our model. Notice how it nearly always visits a major skeleton joint. Figure 6b shows an attention heatmap over all pixels. It illustrates that different regions of the body attract varying levels of attention. Our model easily identifies unique shoes or hair styles. Furthermore, it identifies the left female's hips as a discriminative region. As confirmed in the biomechanics literature [15], females demonstrate strong lateral sway in the hip region. For some females, this alone can be the unique motion signature.

5. Conclusion

We introduced a recurrent attention model that identifies discriminative spatio-temporal regions for the person identification problem from depth video. Our model learns unique volumetric signatures from a high-dimensional 4D input space. Reducing the dimensionality through glimpses and an encoder allows us to train a recurrent network with a LSTM module. Evaluating our model's performance on two, three, and four dimensional inputs showed that our attention model achieves state-of-the-art performance on several person identification datasets. Visualizations of our model's attention offer new insights for future research in computer vision, biomechanics, and physiology.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014.

[3] A. Albiol, J. Oliver, and J. M. Mossi. Who is who at different cameras: people re-identification using depth cameras. *Computer Vision*, 2012.

[4] V. Andersson, R. Dutra, and R. Araújo. Anthropometric and human gait identification using skeleton data from kinect sensor. In *Symposium on Applied Computing*. ACM, 2014.

[5] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *EECV*, 2012.

[6] K. Bashir, T. Xiang, and S. Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 2010.

[7] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*. 2006.

[8] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.

[9] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *ICML*, 2010.

[10] R. J. C. Chen and N. Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *ICIP*, 2015.

[11] J. Caicedo and S. Lazebnik. Semantic guidance of visual attention for localizing objects in scenes. *ICCV*, 2015.

[12] L. Cao, Z. Liu, and T. S. Huang. Cross-dataset action detection. In *CVPR*, 2010.

[13] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals. Listen, attend and spell. *arXiv preprint*, 2015.

[14] X. Chen and C. L. Zitnick. Minds eye: A recurrent visual representation for image caption generation. *CVPR*, 2015.

[15] S. Cho, J. Park, and O. Kwon. Gender differences in three dimensional gait analysis data from 98 healthy korean adults. *Clinical biomechanics*, 2004.

[16] L. Chunli and W. Kejun. A behavior classification based on enhanced gait energy image. In *Networking and Digital Society*. IEEE, 2010.

[17] J. E. Cutting and L. T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society*, 1977.

[18] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas. Learning where to attend with deep architectures for image tracking. *Neural computation*, 2012.

[19] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[20] A. Dubois and F. Charpillet. A gait analysis method based on a depth camera for fall prevention. In *Engineering in Medicine and Biology Society*. IEEE, 2014.

[21] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. *CVPR*, 2015.

[22] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.

[24] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with ensemble localized features. In *ECCV*. 2008.

[25] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. Draw: A recurrent neural network for image generation. *ICML*, 2015.

[26] J. Han and B. Bhanu. Individual recognition using gait energy image. *PAMI*, 2006.

[27] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[28] M. Hofmann, S. Bachmann, and G. Rigoll. 2.5d gait biometrics using the depth gradient histogram energy image. In *Biometrics*. IEEE, 2012.

[29] X. Huang and N. V. Boulgouris. Gait recognition using linear discriminant analysis with artificial walking conditions. In *ICIP*, 2010.

[30] D. Ioannidis, D. Tzovaras, I. G. Damousis, S. Argyropoulos, and K. Moustakas. Gait recognition using compact feature extraction transforms and depth information. *Information Forensics and Security*, 2007.

[31] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[32] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013.

[33] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *PAMI*, 2013.

[34] A. Kale, N. Cuntoor, B. Yegnanarayana, A. Rajagopalan, and R. Chellappa. Gait analysis for human identification. In *Audio-and Video-Based Biometric Person Authentication*. Springer, 2003.

[35] A. Karpathy, J. Johnson, and L. Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint*, 2015.

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

[37] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *PAMI*, 2013.

[38] I. Laptev. On space-time interest points. *IJCV*, 2005.

[39] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, 2007.

[40] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[41] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[42] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[43] M. T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *EMNLP*, 2015.

[44] A. Mansur, Y. Makihara, R. Aqmar, and Y. Yagi. Gait recognition under speed transition. In *CVPR*, 2014.

[45] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *ICANN*. Springer, 2011.

[46] G. Mather and L. Murdoch. Gender discrimination in biological motion displays based on dynamic cues. *Biological Sciences*, 1994.

[47] D. Maturana and S. Scherer. 3d convolutional neural networks for landing zone detection from lidar. In *ICRA*, 2015.

[48] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, 2015.

[49] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.

[50] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014.

[51] A. Mogelmose, T. B. Moeslund, and K. Nasrollahi. Multimodal person re-identification using rgb-d sensors and a transient identification database. In *Biometrics and Forensics*. IEEE, 2013.

[52] M. Munaro, A. Basso, A. Fossati, L. Van Gool, and E. Menegatti. 3d reconstruction of freely moving persons for re-identification with a depth sensor. In *ICRA*, 2014.

[53] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. Van Gool. One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*. Springer, 2014.

[54] M. Munaro, S. Ghidoni, D. T. Dizmen, and E. Menegatti. A feature-based approach to people re-identification using skeleton keypoints. In *ICRA*, 2014.

[55] B. C. Munsell, A. Temlyakov, C. Qu, and S. Wang. Person identification using full-body motion and anthropometric biometrics from kinect videos. In *ECCV*, 2012.

[56] M. P. Murray. Gait as a total pattern of movement: indcluding a bibliography on gait. *American Journal of Physical Medicine & Rehabilitation*, 1967.

[57] M. P. Murray, A. B. Drought, and R. C. Kory. Walking patterns of normal men. *Journal of Bone & Joint Surgery*, 1964.

[58] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, 2013.

[59] F. Pala, R. Satta, G. Fumera, and F. Roli. Multi-modal person re-identification using rgb-d cameras. In *Circuits and Systems for Video Technology*. IEEE, 2015.

[60] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010.

[61] P. Salvagnini, L. Bazzani, M. Cristani, and V. Murino. Person re-identification with a ptz camera: an introductory study. In *ICIP*, 2013.

[62] H. C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *PAMI*, 2013.

[63] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes. Gait energy volumes and frontal gait recognition using depth images. In *Biometrics*. IEEE, 2011.

[64] D. Skog. Gait-based reidentification of people in urban surveillance video. 2010.

[65] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[66] A. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa. Matching shape sequences in video with applications in human movement analysis. *PAMI*, 2005.

[67] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *ECCV*. 2012.

[68] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[69] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.

[70] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. *CVPR*, 2015.

[71] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend, tell: Neural image caption generation with visual attention. *JMLR*, 2015.

[72] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *CVPR*, 2015.

[73] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. *ICPR*, 2014.

[74] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015.

[75] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013.

[76] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*. 2014.

[77] G. Zhao, G. Liu, H. Li, and M. Pietikäinen. 3d gait recognition using multiple cameras. In *Automatic Face and Gesture Recognition*. IEEE, 2006.

[78] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[79] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014.

[80] W. S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *ICCV*, 2015.

[81] M. Zontak, I. Mosseri, and M. Irani. Separating signal from noise using patch recurrence across scales. In *CVPR*, 2013.