The Romanian Academy

Simion Stoilow Institute of Mathematics

Perceptual Models for Computational Visual Analysis

Author,
Elisabeta Marinoiu

Supervisor,
C.S. I Dr. Cristian Sminchișescu

Ph. D. Thesis Summary

Bucharest
2018

Chapter 1

Introduction

When people look at an image or a video, they can immediately understand the underlying 3d space, as well as identify and describe the key actors, actions and objects. Analyzing some of the processing involved in human visual perception may offer insights towards building more robust and adaptable computer vision applications.

This thesis aims at building automatic visual sensing systems and integrate, whenever possible, elements of human visual perception. We firstly address a more general problem, automatic video captioning, i.e. generating a natural-language sentence for describing the content of a video. Then we focus on the perceptual analysis and automatic reconstruction of a more specific but extremely challenging visual category – humans – and their associated 2d and 3d pose. Here we consider only the monocular case, for both images and video sequences. These tasks have applications in fields as diverse as video indexing, behavioral modeling, assisted therapy or self-driving cars.

We propose a video captioning model that leverages spatio-temporal attention mechanisms and recurrent neural networks based on long short-term memory. We integrate additional, potentially valuable, information by relying on spatio-temporal video proposals and classifier responses for semantic categories. The task is challenging due to the large variety of semantic categories and textual annotations. Our method produces competitive, state-of-the art results, while localizing semantic concepts (subject, verbs, objects) with no additional supervision, over space and time.

Next, we present an experimental apparatus that aims to link human perception

with quantitative measurements. We asked subjects to re-enact the 3d pose seen in a single person image, following a short exposure time. We precisely tracked their body movements and recorded their eye fixations using specialized equipment. We provide an extensive analysis of eye movement consistency as well as quantitative and qualitative performance of the pose re-enactment. Our data and insights are further used to learn perceptual metrics that produce more stable and meaningful results, when integrated with automated single person 3d pose estimation predictors.

Lastly, we extend single person 3d pose predictors towards an automated multiple people sensing system that estimates the 2d, 3d pose and shape, as well as camera scene translation of multiple people in natural images. The main challenges arise from the variety of the human bodies, occlusions and partial views, as well, the close interactions, and the ambiguity of monocular perspective projection. We leverage a deep multi-task sensing network with an optimization step guided by detailed semantic cues. We enforce a series of scene constraints, *e.g.*, ground plane support and simultaneous volume occupancy exclusion, that lead to state-of-the art results as well as promising qualitative reconstructions in natural images.

Chapter 2

Spatio-Temporal Attention Models for Grounded Video Captioning

Automatic video captioning is challenging due to the complex interactions in dynamic real scenes. A comprehensive system would ultimately localize and track the objects, actions and interactions present in a video and generate a description that relies on temporal localization in order to ground the visual concepts. However, most existing automatic video captioning systems map from raw video data to high level textual description, bypassing localization and recognition, thus discarding potentially valuable information for content localization and generalization. In this work we present an automatic video captioning model that combines spatio-temporal attention and image classification by means of deep neural network structures based on long short-term memory. The resulting system is demonstrated to produce state-of-the-art results in the standard YouTube captioning benchmark while also offering the advantage of localizing the visual concepts (subjects, verbs, objects), with no grounding supervision, over space and time.

2.1 Methodology

Our approach to video captioning has two main components: first revealing the spatio-temporal visual support of words in video and then guiding the sentence generation

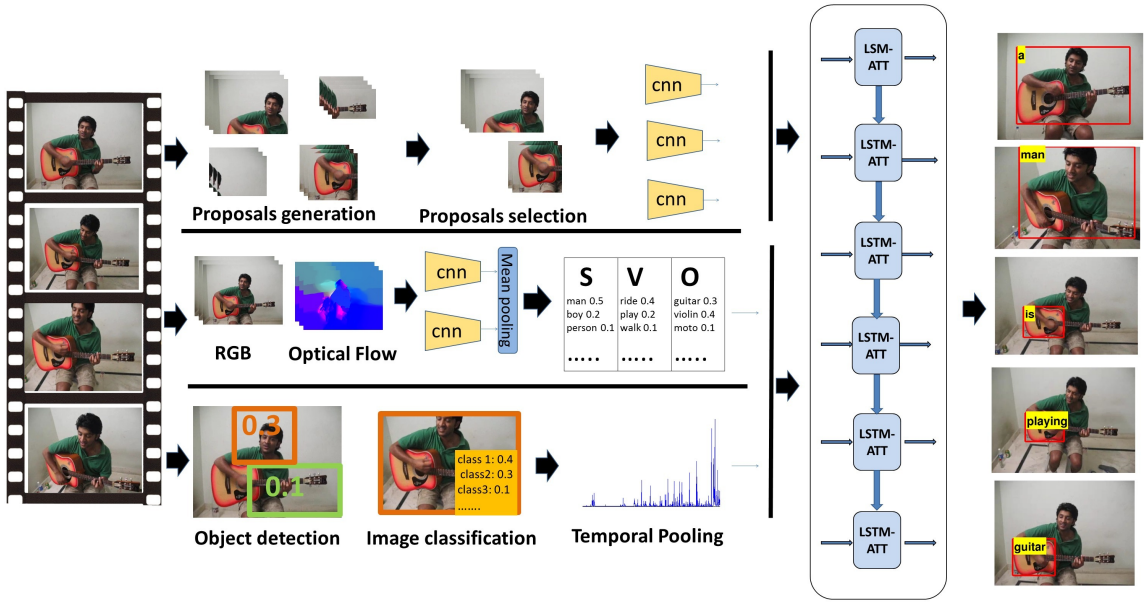


Figure 2-1: Overview of our approach for aitomatic video captioning and spatio-temporal grounding of semantic concepts.

process by including semantic information in the learning process. We integrate a soft-attention mechanism, operating over a pool of spatio-temporal proposals, into a state-of-the-art recurrent network. The joint model learns to produce semantically meaningful sentences while attending to different parts of the video. The semantic information is obtained in two ways: (a) by learning to predict subjects, verbs and objects (S,V,O) and (b) by using pre-trained state-of-the-art image classification and object detection models. An overview of our modeling and computational pipeline is shown in figure 2-1.

2.1.1 Spatio-Temporal Object Proposals

We use the method from [15] to gather a pool of spatio-temporal object proposals. We split each video into parts using a shot boundary detection method [12]. Around 1,000 spatio-temporal proposals are extracted separately for each sub-video and together they form the pool of proposals for the whole video. These are then filtered and ranked according to a series of heuristics that take into account their spatial and temporal extent as well as the probability that they contain an object. We represent a video by

m such descriptors corresponding to best scoring m spatio-temporal proposals. Given a proposal, for each of its bounding boxes in the video frames, we extract the output of the fc_7 layer of the VGG-19. The feature descriptor for a proposal is obtained by mean-pooling over all the bounding boxes.

2.1.2 Attention-based LSTM

Soft-Attention Mechanism. We incorporate a soft-attention mechanism into the LSTM in order to allow the model to selectively focus on different parts of the video, represented by a series of spatio-temporal object proposals, each time it produces a word. The Attention-based LSTM learns to weight the proposals every time a word is produced, thus being able to indicate what is the *localized* visual support used to produce a particular word from the video description.

2.1.3 High-Level Semantic Description

For improved generalization of our model, we incorporate several types of semantic information: image classifiers and object detectors responses as well as responses from classifiers for subjects, verbs and objects that we learn ourselves. Generating a textual description of a video requires identifying the actors and their interactions and then constructing a grammatically well-formed sentence. For this purpose, in order to generate a human-like textual description of a video, we first represent a video in the form of a Subject(S), a Verb(V) and an Object(O) (similarly to earlier works [7]). We then integrate this representation with state-of-the-art recurrent models, along with spatio-temporal localization processes and object detection and classification information. In order to learn a semantic high-level representation for each video, we represent a sentence in a compact and simplified manner that preserves its main idea by extracting a (S,V,O) tuple - e.g. the sentence *A cat plays with a toy* is represented as (cat, play, toy)). We treat the three vocabularies separately and use Least Squares Support Vector Machine (LS-SVM) as a classifier in a one-vs-all approach.

2.2 Experimental Details

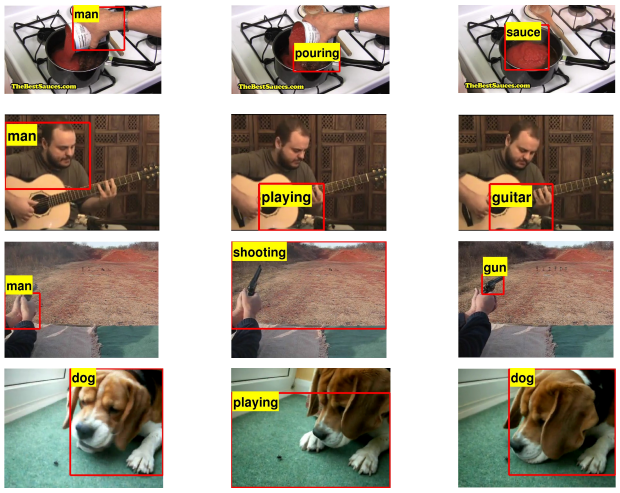
Dataset Description. We perform our experiments on the YouTube dataset [4] which consists of 1,967 short videos (between 10s and 25s length) collected from YouTube that usually depict only one main activity. Each video has approximately 40 human-generated English descriptions collected through Amazon Mechanical Turk.

Evaluation Measures. We report our results under BLEU [18] and METEOR [11] metrics which were originally proposed for the evaluation of automatic translation approaches and have also been adopted by previous works in video and image captioning.

2.2.1 Experimental Results

Quantitative Results. Results obtained with the proposed models are shown in table 2.1. Our attention-based recurrent neural network model (LSTM-ATT) model achieves competitive results compared to other methods. Adding semantic features on top of this model improves the state-of-the-art results on the BLEU@n metric, while also performing well on METEOR. The contributions of the SVO semantic features alone and in conjunction with additional DET (detection) and CLS (classification) features are also presented. In the case of SVO features alone, the best results are obtained with LSTM2-ATT(SVO) method for both evaluation metrics (BLEU@4 52.0%, METEOR 32.3%), while when using the full semantic features, our best performing method under BLEU is LSTM-ATT(SVO,DET,CLS) (50.6%) and under METEOR is LSTM2-ATT(SVO,DET,CLS) (32.4%).

Qualitative Results. Our attention mechanism, built on top of spatio-temporal object proposals, allows for a *visual explanation* of what the model ranked as the most relevant visual support for emitting a particular word. This can be done by inspecting the learned weights and their associated proposals. In figure 2-2 we show the proposal with the highest associated weight that was used in generating a particular word.



Ours: A **man** is **pouring** some **sauce**.
Ref: A man is pouring some sauce into a pan.

Ours: A **man** is **playing guitar**.
Ref: A man is playing a guitar.

Ours: A **man** is **shooting** with a **gun**.
Ref: A guy is shooting a gun.

Ours: A **dog** is **playing** with a **dog**.
Ref: A dog is playing with a fly.

Figure 2-2: Highest scoring proposals of our model for each emitted word in the sentence. We only illustrate the grounding of the main words in the sentence and ignore linking words. The complete sentence is shown in the right column together with the closest reference from the human annotations. For each proposal we show a single, randomly selected frame.

Method	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR
FGM [20]	-	-	-	13.68	23.9
S2VT[21]	-	-	-	-	29.8
MM-VDN[23]	-	-	-	37.64	29.00
LSTM-YT-coco [22]	-	-	-	33.29	29.07
LSTM-YT-coco+flicker [22]	-	-	-	33.29	28.88
Temporal attention [24]	-	-	-	41.92	29.60
LSTM-E (VGG+C3D) [17]	78.8	66.0	55.4	45.3	31.0
h-RNN[25]	81.5	70.4	60.4	49.9	32.6
HRNE with attention[16]	79.2	66.3	55.1	43.8	**33.1**
GRU-RCNN [1]	-	-	-	49.63	31.70
LSTM	78.0	66.4	56.7	45.4	31.2
LSTM(SVO)	80.1	68.1	57.5	45.8	31.2
LSTM(DET,CLS)	81.2	68.9	57.9	46.2	31.1
LSTM(SVO,DET,CLS)	80.8	69.3	59.3	48.3	30.7
LSTM-ATT	80.1	68.9	59.4	48.7	31.9
LSTM-ATT(SVO)	81.0	70.5	61.2	50.5	32.3
LSTM-ATT(DET,CLS)	81.9	70.9	60.9	50.5	31.6
LSTM-ATT(SVO,DET,CLS)	82.0	71.6	62.4	51.5	32.0
LSTM2-ATT(SVO)	**82.4**	**71.8**	**62.5**	**52.0**	32.3
LSTM2-ATT(DET,CLS)	80.6	68.1	57.4	46.0	31.8
LSTM2-ATT(SVO,DET,CLS)	81.5	70.8	61.5	50.6	32.4

Table 2.1: Comparison with previous works on BLEU@1 - BLEU@4 and METEOR metrics. Values are reported as percentage %.

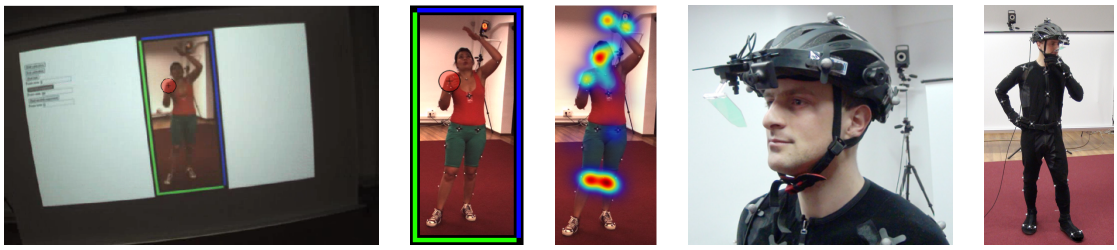
Chapter 3

Pictorial Human Spaces: A Computational Study on the Human Perception of 3d Articulated Poses

Human motion analysis in images and video, with its deeply inter-related 2d and 3d inference components, is a central computer vision problem. Yet, there are no studies that reveal how humans perceive other people in images and how accurate they are. In this chapter we aim to unveil some of the processing–as well as the levels of accuracy–involved in the 3d perception of people from images by assessing the human performance. Moreover, we reveal the quantitative and qualitative differences between human and computer performance when presented with the same visual stimuli and show that metrics incorporating human perception can produce more meaningful results when integrated into automatic pose prediction algorithms.

3.1 Apparatus for Human Pose Perception

We propose an experimental apparatus that allows linking a partially subjective phenomenon like the 3d human pose perception with measurement. Our approach is to



(a) (b) (c) (d) (e)

Figure 3-1: Illustration of our human pose perception apparatus. (a) Screen on which the image is projected as captured by the external camera of the eye tracker. (b) Result of mapping the fixation distribution on the original high-resolution image, following border detection, tracking and alignment. (c) Heat map distribution of all fixations of one of our subjects for this particular pose. (d) Detail of our head-mounted eye tracker and (e) 3d motion capture setup.

dress people in a motion capture suit, equip them with an eye tracker and show them images of other people in different poses, which were obtained using motion capture as well (fig. 3-1). By asking the subjects to re-enact the poses shown, we can link perception and measurement. We use a state-of-the-art Vicon motion capture system together with a head mounted, high-resolution mobile eye tracking system.

3.1.1 Experimental Design and Dataset Collection

Subjects and General Setup. We first analyze the re-enactment performance of 10 subjects, 5 male and 5 female, who did not have a medical history of eye problems or mobility impediments. Moreover, their profession did not require above average neuro-motor skills (as required in the case of dancing, acting or practicing a particular sport). We will refer to this group as the *regular* subjects. We also analyze the performance of another 4 subjects, 2 males and 2 females, who were all final year choreography students, focusing on modern and classical ballet. This group will be referred to as the *skilled* subjects. The images were projected on a 1.2 meters tall screen located 2.5–3 meters away.

Each subject was required to stand still and look at one image at a time until it disappeared, then re-enact the pose by taking as much time as necessary. For each

pose projected on the screen, we captured both the scanpaths and the 3d movement of the subject in the process of re-enacting the pose, once it had disappeared from the screen. Once the 5s exposure time has passed, the subject no longer had the possibility to see the image of the pose to re-enact, but had to adjust his position based on the memory of that pose. We display a total of 120 images, each representing a bounding box of a person. The images are mainly frontal. 100 contain easily reproducible standing poses, whereas 20 of them are harder to re-enact as they require sitting on the floor, which often results in self-occlusion. The poses shown were selected from Human3.6M [9], from various types of daily activities.

3.2 Data Analysis

3.2.1 Human Eye Movement Recordings

Static and Dynamic Consistency. In this section we analyze how consistent the subjects are in terms of their fixated image locations. We are first concerned with evaluating static consistency, which considers only the fixation locations and then with dynamic consistency, which takes into account the order of fixations. To evaluate how well the subjects agree on fixated image locations, we predict each subject's fixations in turn using the information from the other subjects [6, 14]. This was done considering the same pose as well as different poses. Fig. 3-2 indicates good consistency.

To evaluate how consistent the subjects are in their order of fixating areas of interest (AOIs), we used the hidden Markov modeling recently developed by [14]. The states correspond to AOIs that were fixated by subjects and the transitions correspond to saccades. For each pose, we learn a dynamic model from the scanpaths of 9 subjects and compute the likelihood of the 10^{th} subject's scanpath under the trained model. The leave-one-out process is repeated in turn for each subject and the likelihoods are averaged. The average likelihood (normalized by the scanpath length) obtained is -9.38. Results are compared against the likelihood of randomly generated

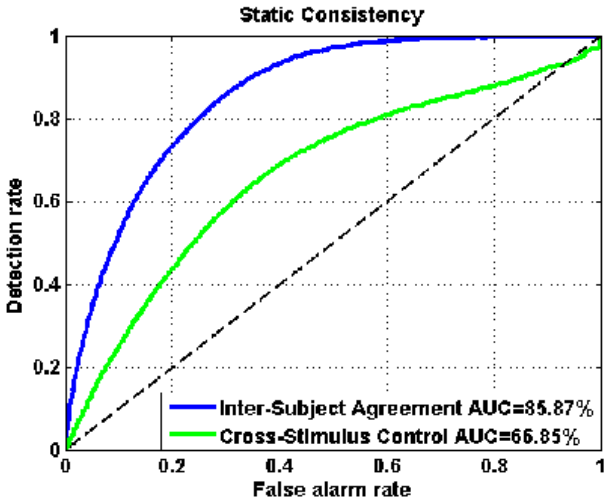


Figure 3-2: Static inter-subject eye movement agreement. Fixations from one subject are predicted using data from the other 9 subjects both on the same image (blue) and on a different image of a person, randomly selected from our 120 poses (green).

scanpaths and the likelihoods of scanpaths from another randomly chosen pose.The average likelihood is much smaller for randomly generated trajectories (-42.03) than for those of human subjects. Also, the likelihood of scanpaths obtained from other images is considerably smaller -17.12 than the likelihood of scanpaths obtained from the same image indicating that subjects are consistent in the order they fixate AOIs.

Which are the most fixated joints? We study whether certain joints are fixated more than others and we want to know whether this would happen regardless of the pose shown, or whether it varies with the pose. For this purpose, we consider the number of fixations that fall on a particular joint. Fig. 3-3 shows the distribution of fixations on body joints, averaged over poses. Notice that the wrists and the head area are the most looked at, within a general trend of fixating upper body parts more than lower ones.

3.2.2 3D Pose Re-Enactment

In this section we complement eye-movement studies with an analysis of how well humans are able to reproduce the 3d poses of people shown in images.

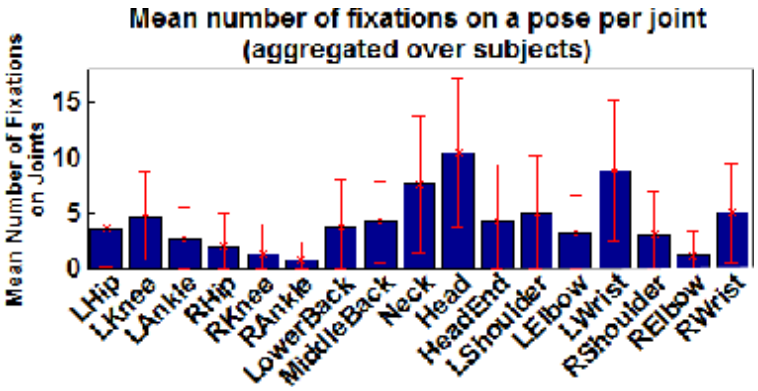


Figure 3-3: Fixation counts on each joint. The mean and standard deviation is computed among the 120 poses by aggregating over all 10 subjects, for each pose.

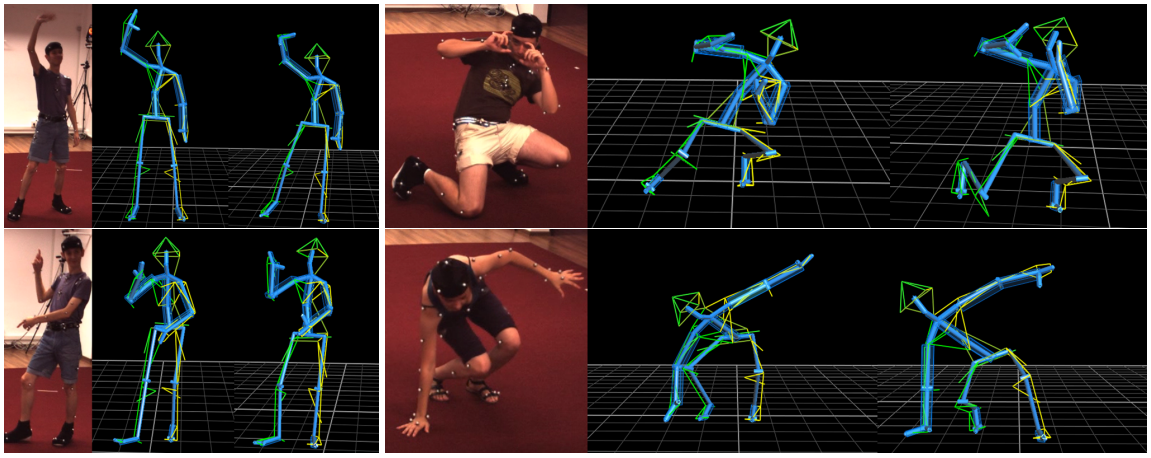


Figure 3-4: Examples of subject re-enactment for two easy (left) and two hard (right) poses. For each pose, the first re-enactment shown is from a regular subject, whereas the second one is from a a skilled one.

How accurately do humans re-enact 3d poses?

We have compared the re-enactment performance of regular and skilled subjects on the same stimuli. We want to understand to what extent formal training in professions requiring sharp neuro-motor skills and good body positioning self-awareness influences perception and the capacity to re-enact poses, as measured under the widely used metric, MPJPE. Table 3.1 shows re-enactment errors under the MPJPE for regular and skilled subjects, respectively. It can be noticed that the overall completion error of the skilled subjects is only 7.1 mm (under MPJPE) smaller that the overall completion error of the regular subjects. The small difference in errors between the two groups

of subjects suggests that: (1) the metrics used are not sensitive enough to the criteria optimized by people perceiving and re-enacting 3D poses, (2) having superior mobility and body coordination does not make a significant difference in the context of the task analyzed here. Examples of re-enactment from both skilled and regular subjects on 2 easy and 2 hard poses are presented in fig. 3-4.

	MPJPE error (mm)		
Method	Easy	Hard	Both
Regular Subject	91.7 ± 35.9	156.7 ± 58.1	102.5 ± 58.1
Skilled Subject	83.9 ± 31.6	153.1 ± 65.6	95.4 ± 47.3
KDE	100.33 ± 39.54	267.42 ± 133.22	128.18 ± 89.45

Table 3.1: Subject re-enactment results as well as KDE prediction for easy poses, hard poses and over all poses under the MPJPE metric.

3.3 Perceptual Metrics for Automatic 3d Human Pose Estimation

In this section we focus on two aspects: 1) understanding the types of errors humans make in pose re-enactment and compare them with those of automatically generated poses, 2) learning perceptual metrics that more truthfully reflect the semantics of a human pose.

3.3.1 Human vs Computer Vision Performance

We aim to reveal the quantitative and qualitative differences between poses re-enacted by humans and poses estimated by a computer vision model and algorithm when presented with the *same* visual stimuli. We consider a structured prediction model, Kernel Dependency Estimation (KDE) [5], to learn a mapping from features extracted from the image of a person to the 3d joint representation of his/her pose. For training we use Human80K which is a subset of the larger Human3.6m, from which we selected the 120 poses shown to subjects for re-enactment.

What is the error difference between humans and a vision model (KDE)? In table 3.1 we show the average re-enactment error for both skilled and regular subjects and the average KDE prediction error under MPJPE metric. On easy poses, the difference between skilled subjects and KDE predictions is 16mm while between regular subjects and KDE predictions is 9mm. On the hard poses, however, the difference between subjects and KDE is significantly higher: 110mm in the case of regular subjects and 113mm in the case of skilled subjects. This indicates that although both human and algorithmic performances are diminished when presented with hard poses, the algorithmic approach struggles considerably more than humans when the poses are mainly seated and have severe self-occlusions.

3.3.2 Perceptual Metric Learning

In this section we present our proposal of learning a new metric that captures the perceptual difference between poses. In this way, we aim to reduce the gap between the human perception of pose similarity and what the commonly used metric (MPJPE) evaluates. For this purpose we use the re-enactments of both skilled and regular subjects to learn a perceptual metric over poses. To learn a perceptually relevant metric from subject re-enactment, we use Relevant Component Analysis (RCA) [2] which changes the feature space by a global linear transformation. It assigns high weights to 'relevant dimensions' and low weights to 'irrelevant dimensions'.

3.3.3 Perceptual Metric in Pose Estimation

In this section we integrate the newly learned perceptual metric in the KDE framework. We train the model on Human80K [8] dataset using both a Gaussian kernel $K_Y(x, y) = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}}$, and a perceptual kernel $K_Y(x, y) = e^{-\frac{d_P^2(x,y)}{2\sigma^2}}$ on the target variables, where $d_P(x, y)$ is the perceptual metric between poses x and y learned as described in §3.3.2.

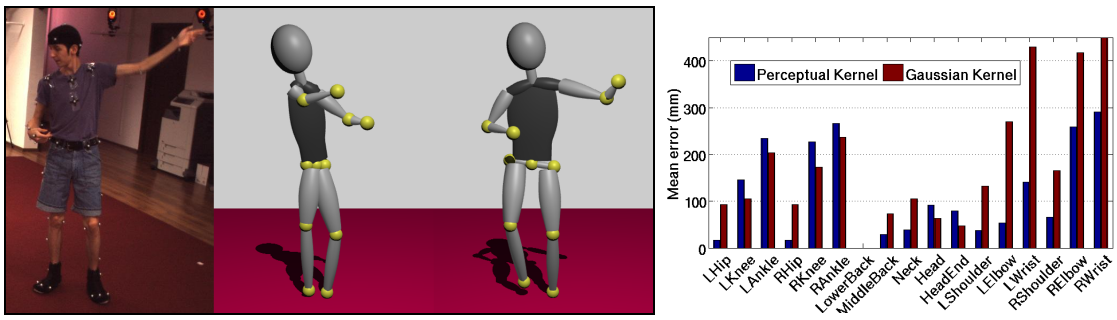
In table 3.2 we show the mean MPJPE and mean perceptual error for poses estimated with a Gaussian kernel and a perceptual kernel, respectively. It can be

noticed that those produced by the perceptual kernel are smaller under both metrics.

Method	Mean MPJPE	Mean Perceptual Error
KDE - Gaussian	113.07 ± 72	18.44 ± 10.25
KDE - Perceptual	109.16 ± 66	14.92 ± 9.43

Table 3.2: Pose estimation errors under metric based on both an Euclidean and a perceptual kernel.

Poses with similar MPJPE error can look perceptually very different if a subset of the joints have very large errors. In figure 3-5 we show an example when the computer vision model prediction obtained using the perceptual kernel, although not perfect, appears qualitatively better than the one obtained using the Gaussian kernel. We also show the distribution of error per joints for the two predictions. Notice that even if most of the joints in the two predictions have similar errors, there are 4 joints (left elbow, left wrist, right elbow, right wrist) with extremely large errors, making the pose perceptually very different from the one shown in the stimulus image.



(a) (b) (c) (d)

Figure 3-5: a) Test image, b) Prediction of KDE with Gaussian kernel, c) Prediction of KDE with perceptual kernel, d) Per joint error distribution for the prediction obtained with the Gaussian and perceptual kernel, respectively.

Chapter 4

Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes *The Importance of Multiple Scene Constraints*

Human sensing has greatly benefited from recent advances in deep learning, parametric human modeling, and large scale 2d and 3d datasets. However, existing 3d models make strong assumptions about the scene, considering either a single person per image, full views of the person, a simple background or many cameras. In this chapter, we leverage state-of-the-art deep multi-task neural networks and parametric human and scene modeling.. We perform experiments on both single and multi-person datasets, and systematically evaluate each component of the model, showing improved performance and extensive multiple human sensing capability. We also apply our method to images with multiple people, severe occlusions and diverse backgrounds captured in challenging natural scenes, and obtain results of good perceptual quality.

4.1 Multiple Persons in the Scene Model

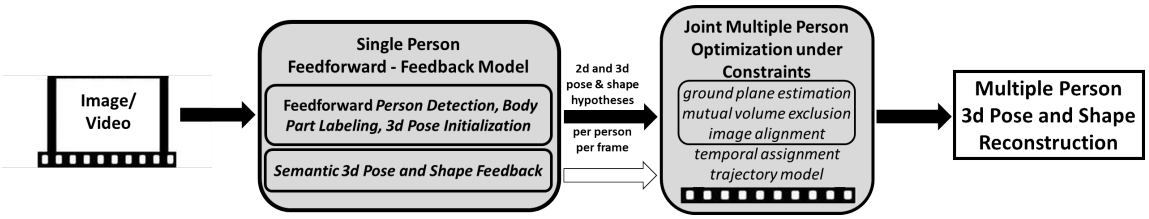


Figure 4-1: **Processing pipeline** of our monocular model for the estimation of 3d pose and body shape of multiple people. The system combines a single person model that incorporates feedforward initialization and semantic feedback, with additional constraints such as ground plane estimation, mutual volume exclusion, and joint inference for all people in the scene. For monocular video, the 3d temporal assignment of people is resolved using a Hungarian method, and trajectory optimization is performed jointly over all people and timesteps, under all constraints, including image consistency, for optimal results.

Problem formulation. Without loss of generality, we consider N_p uniquely detected persons in a video with N_f frames. Our objective is to infer the best pose state variables $\mathbf{\Theta} = [\boldsymbol{\theta}_p^f] \in \mathbb{R}^{N_p \times N_f \times 72}$, shape parameters $\mathbf{B} = [\boldsymbol{\beta}_p^f] \in \mathbb{R}^{N_p \times N_f \times 10}$ and individual person translations $\mathbf{T} = [\mathbf{t}_p^f] \in \mathbb{R}^{N_p \times N_f \times 3}$, with $p \in N_p$ and $f \in N_f$. We start by first writing a per-frame, person-centric objective function $L_I^{p,f}(\mathbf{B}, \mathbf{\Theta}, \mathbf{T})$

$$L_I^{p,f} = L_S^{p,f} + L_G^{p,f} + L_R^{p,f} + \sum_{\substack{p'=1 \\ p' \neq p}}^{N_p} L_C^f(p, p'), \tag{4.1}$$

where the cost L_S takes into account the visual evidence computed in every frame in the form of semantic body part labeling, L_C penalizes simultaneous (3d) volume occupancy between different people in the scene, and L_G incorporates the constraint that some of the people in the scene may have a common supporting plane. The term $L_R^{p,f} = L_R^{p,f}(\boldsymbol{\theta})$ is a Gaussian mixture prior similar to [3]. The image cost for multiple people under all constraints can be written as

$$L_I^f = \sum_{p=1}^{N_p} L_I^{p,f} \tag{4.2}$$

If a monocular video is available, the static cost L^f is augmented with a trajectory model applicable to each person once the temporal assignment throughout the entire video has been resolved. The complete video loss writes

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_T = \sum_{p=1}^{N_p} \sum_{f=1}^{N_f} \left(L_I^{p,f} + L_T^{p,f} \right) \tag{4.3}$$

where L_T can incorporate prior knowledge on human motion, ranging from smoothness, assumptions of constant velocity or acceleration, or more sophisticated models learned from human motion capture data.

In order to infer the pose and 3d position of multiple people we rely on a parametric human representation, SMPL [13], with a state-of-the-art deep multitask neural network for human sensing, DMHS [19]. In practice, we cannot assume a constant number of people throughout a video and we first infer the parameters $\mathbf{B}$,$\mathbf{\Theta}$, $\mathbf{T}$ independently for each frame by minimizing the sum of the first two cost functions: L_S and L_C. Then, we temporally track the persons obtained in each frame by means of optimally solving an assignment problem, then re-optimize the objective, by adding the temporal and ground plane constraints, L_T and L_G. An overview of the method is shown in fig. 4-1.

4.2 Experiments

We numerically test our inference method on two datasets, CMU Panoptic [10] and Human3.6M [9], as well as qualitatively on challenging natural scenes (see fig. 4-2). Given a video with multiple people, we first detect the persons in each frame and obtain initial feedforward DMHS estimates for their 2d body joints, semantic segmentation and 3d pose.

Human3.6M is a large-scale dataset that contains single person images recorded in a laboratory setup using a motion capture system. We select 3 of the most difficult actions: *sitting*, *sitting down* and *walking dog* to test our single-person model. We use the official left-out test set from the selected actions, consisting of 160K examples.

	Haggling		Mafia		Ultimatum		Pizza		Mean	
Method	Pose	Translation	Pose	Translation	Pose	Translation	Pose	Translation	Pose	Translation
DMHS [19]	217.9	-	187.3	-	193.6	-	221.3	-	203.4	-
2d Loss	**135.1**	282.3	174.5	502.2	**143.6**	357.6	177.8	419.3	157.7	390.3
Semantic Loss	144.3	260.5	179.0	459.8	160.7	376.6	178.6	413.6	165.6	377.6
Smoothing	141.4	260.3	173.6	454.9	155.2	368.0	173.1	403.0	160.8	371.7
Smoothing Ground Plane	140.0	**257.8**	**165.9**	**409.5**	150.7	**301.1**	**156.0**	**294.0**	**153.4**	**315.5**

Table 4.1: Automatic 3d human pose and translation estimation errors (in mm) on the Panoptic dataset (9,600 frames, 21,404 people). Notice the value of each component and the impact of the ground-plane constraint on correct translation estimation.

Method	WalkingDog	Sitting	Sitting Down
DMHS [19]	78	119	106
Semantic Loss	75	109	101
Multi View	51	71	65
Smoothing	**48**	**68**	**64**

Table 4.2: Mean per joint 3d position error (in mm) on the Human3.6M dataset, *evaluated on the test set* of several very challenging actions. Notice the importance of various constraints in improving estimation error.

We show results in table 4.2. We obtain an improvement over DMHS by using the proposed semantic 3d pose and shape feedback.

CMU Panoptic Dataset. We selected data from 4 activities (*Haggling*, *Mafia*, *Ultimatum* and *Pizza*) which contain multiple people interacting with each other. In total, we obtain 9,600 frames that contain 21,404 people. We do not validate/train any part of our method on this data.

Evaluation Procedure. We evaluate both the inferred pose, centered in its hip joint, under mean per joint position error (MPJPE), and the estimated translation for each person under standard Euclidean distance. We perform the evaluation for each frame in a sequence, and average the results across persons and frames.

Ablation Studies. We systematically test the main components of the proposed monocular inference system and show the results detailed for each activity in table 4.1. Compared to DMHS, our complete method reduces the MPJPE error significantly, from 203.4 mm to 153.4 mm on average (-25%), while also computing the translation of each person in the scene. The translation error is, on average, 315.5

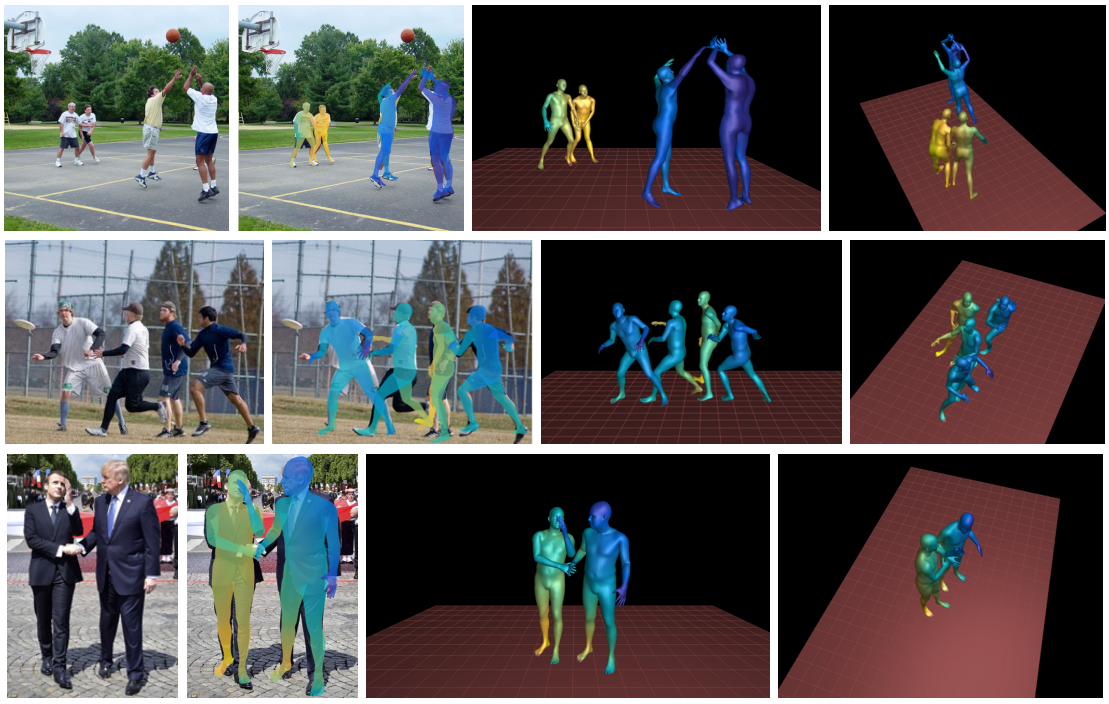


Figure 4-2: **Automatic 3d reconstruction of multiple people from monocular images of complex natural scenes**. Left to right: input image, inferred model overlaid, and two different views of 3d reconstructions obtained by our model (including ground plane). Challenging poses, occlusions, different scales and close interactions are correctly resolved in the reconstruction.

mm. The semantic projection term helps disambiguate the 3d position of persons and reduces the translation error compared to using only the 2d projection term. Temporally smoothing the pose estimates decreases the translation error further. Imposing the ground plane constraint makes the most significant contribution in this setup, decreasing the total translation error from 371 mm to 315 mm (-15%). Our method produces perceptually plausible 3d reconstructions with good image alignment in scenes with many people, some only partially visible, and captured under non-conventional viewing angles.

We have presented a monocular model for the integrated 2d and 3d pose and shape estimation of multiple people, under multiple scene constraints.

Bibliography

[1] N. Ballas, L. Yao, C. Pal, and A. C. Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015.

[2] A. Bar-hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *International Conference on Machine Learning*, 2003.

[3] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

[4] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, June 2011.

[5] C. Cortes, M. Mohri, and J. Weston. A general regression technique for learning transductions. In *International Conference on Machine Learning*, pages 153–160, 2005.

[6] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 2009.

[7] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[8] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1661–1668, 2014.

[9] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. Submitted.

[10] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015.

[11] A. Lavie and A. Agarwal. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[12] R. W. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Electronic Imaging'99*, pages 290–301. International Society for Optics and Photonics, 1998.

[13] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH*, 34(6):248:1–16, 2015.

[14] S. Mathe and C. Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in Neural Information Processing Systems*, 2013.

[15] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection proposals. In *ECCV*, 2014.

[16] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. *arXiv preprint arXiv:1511.03476*, 2015.

[17] Y. Pan, T. M. , T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *arXiv preprint arXiv:1505.01861*, 2015.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002.

[19] A. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *CVPR*, 2017.

[20] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, August 2014.

[21] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence – video to text. In *ICCV*, 2015.

[22] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *NAACL HLT*, 2015.

[23] H. Xu, S. Venugopalan, V. Ramanishka, M. Rohrbach, and K. Saenko. A multi-scale multiple instance video description network. In *arXiv preprint arXiv:1505.05914*, 2015.

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

[25] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, June 2016.