# Human Pose Estimation in Video via MCMC Sampling

Evgeny Shalnov and Anton Konushin

*Department of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia*

Keywords: Computer Vision, Human Pose Estimation, MCMC, Kalman Filter.

Abstract: We describe a method for the human pose estimation in a video sequence. We propose a new mathematical model of a human pose in a video sequence, which incorporates motion and pose parameters. We show that the model of (Park and Ramanan, 2011) is a particular case of our model. We introduce a framework to infer an approximation of the optimal value in the proposed model. We use an exact algorithm of motion parameters estimation to reduce complexity of inference. Our approach outperforms results of (Park and Ramanan, 2011) in the most complicated video sequences.

## 1 INTRODUCTION

Video analysis is an extremely important task in computer vision and machine learning. It means a construction of a high-level video description, that can include:

- description of the scene geometry;
- description of people in the video sequence;
- person parameters in each video frame including location and pose;

High-level video description has a lot of potential applications. The security surveillance is one such example. Moreover, the results of video analysis can be applied for the efficient compression based on high-level representation of the input video.

In the paper we focus on the Human Pose Estimation (HPE). We are interested in an estimation of a human pose in a whole video sequence jointly. An accurate description of a human pose in the input video sequence makes it possible to reduce complexity of the estimation of such global person attributes as a physique and a color of clothes.

The lack of efficient and highly accurate techniques of video analysis significantly reduces practical usage of video surveillance systems. Let us assume that it is required to find in the input video sequence a dark-haired man in yellow T-shirt and blue trousers. In average a human operator had to watch a half of the input video sequence to find such person. An automatic technique of a person description construction would significantly reduce complexity of this problem. For an accurate person description it is insufficient to have only approximate location from object tracking. Head, body and limbs should be localized as well.

We propose a new method for human pose estimation in video sequence. The main contributions of our work are:

- We expand mathematical model of the human pose in a video with the hidden parameters. That parameters describe motion of the observed human.
- We show that the basic model of (Park and Ramanan, 2011) is a particular case of ours.
- We convert the problem of an optimal hidden state estimation to an inference in a Linear Dynamical System (LDS).
- We introduce a framework for human pose estimation in a video based on MCMC sampling technique. The proposed framework allows approximate inference of both local (depends on a single frame and its direct neighbors) and global parameters (depends on all frames of the input video sequence).

## 2 HPE VIA SAMPLING

### 2.1 Task Definition

Two main approaches to human pose definition exist. The traditional approach defines human pose in terms of a set of human body parts (fig. 1, a). A head, a

Shalnov E. and Konushin A..
Human Pose Estimation in Video via MCMC Sampling.
DOI: 10.5220/0005462000710079
In *Proceedings of the 5th International Workshop on Image Mining. Theory and Applications* (IMTA-5-2015), pages 71-79
ISBN: 978-989-758-094-9

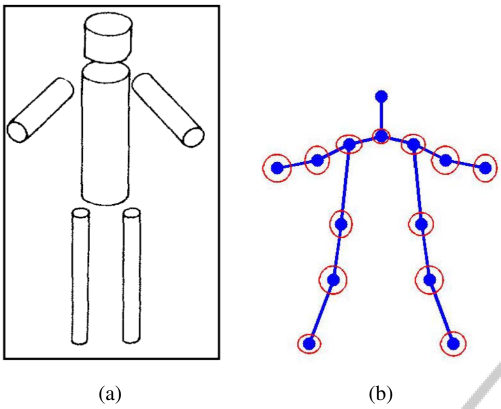


Figure 1: Pose models in a still frame. (a) is the classic articulated limb model of (Marr and Nishihara, 1978), (b) is the model of (Yang and Ramanan, 2011).

body and a thigh are several examples of such parts. Modern models defines a human pose in tersm of a set of human body joints (fig. 1, b). A shoulder, a knee and an elbow are some examples of such joints. This definitions are equivalent. Indeed, the location of each part is uniquely defined by locations of the corresponded joints, and a location of a joint is uniquely defined by a location of the corresponded human body part. In the research we exploit the second definition as it makes the inference easier (Yang and Ramanan, 2011).

Most of the previous approaches to the human pose estimation task work with still frames only. In such case, the problem is usually reduced to inference in a tree structured graphical model. In this models vertices correspond to locations of the joints and edges define limitations on relative joint locations (Yang and Ramanan, 2011). Due to the structure of the graphical model the best configuration of a joint location in a still frame can be obtained efficiently. In spite of significant progress in techniques of the human pose estimation in a still frame, their accuracy is far from ideal. Therefore we propose to improve the accuracy of pose estimation by considering all video frames simultaneously. The proposed approach uses evidence of the pose from the other frames for the result gaining in the current frame. We use the tracking approach (Shalnov and Konushin, 2013) to initially estimate trajectory of the person. A trajectory means an approximate location of the person in each frame of the input video sequence. Hence, the formal input of our algorithm consists of:

- video sequence $W = \{I_t\}$;
- trajectory of the person $Ba = \{B_t\}$.

The output of the algorithm is:

- human pose in the input video sequence. It means location of joints and a scale parameter of the person in each frame of the input video $Pa = \{P_t\}$.

## 2.2 Basic Model

Our research was inspired by the work of (Park and Ramanan, 2011) on human pose estimation in video sequence. They use the mathematical model of human pose in a still frame (Yang and Ramanan, 2011) and expand the inference algorithm. Compared to the previous method the extension by (Park and Ramanan, 2011) allows inference N-best configurations from the model, ensuring that they do not overlap according to some user-provided definition of overlap.

Moreover, they include a simple temporal context from neighboring frames in the model. It allows them to select better pose hypothesis in each frame of the input video sequence. This way they converts the problem of the human pose estimation in video to the following maximization task:

$$\begin{aligned} Pa^* &= argmax\ Score(Pa) \\ Score(Pa) &= \sum_t \Phi(P_t) + \alpha \sum_t \Psi(P_t, P_{t-1}) \end{aligned} \quad (1)$$

where $\Phi(P_t)$ is the score of candidate pose $P_t$ computed by the proposed detector, and $\Psi(P_t, P_{t-1})$ is the (negative of the) total squared pixel difference between each joint in pose $P_{t-1}$ and pose $P_t$.

A set of available inference algorithms is the key distinction between human pose models in still frame and in video. A dynamic programming algorithm is usually applied to infer optimal pose in a still frame. However, it cannot be utilized to infer the optimal set of poses in video. Indeed, the poses in instant of time $t_1$ and $t_2$ are conditionally independent given a pose at instant of time $t \in [t_1, t_2]$ at least. Therefore, inference with the dynamic programming algorithm requires $O(L^{2K})$ elements stored in memory, where L is a number of possible joint locations in the frame and K is a number of joints in the model. The authors use an approximate algorithm. They restrict the possible poses in each frame with best hypotheses. It makes the dynamic programming tractable.

## 2.3 Proposed Model

We use the same model of human pose in a still frame of (Yang and Ramanan, 2011), but with different temporal context. The temporal context of the original model requires the shift of joint location between subsequent frames to be small. In practice, this constraint is a poor motion model for a majority of body joints. For example, the Brownian movement and the constant motion with the same velocity have equal impact to the objective function.

Therefore, we propose to use another temporal context. Our temporal context prefers constant motion of the joints. We are aware of the model doesn't fully corresponds to the joint motion models observed in practice. For example, a knee has a periodical motion model during walking. This periodicity is specified by the a cyclicity of a step. We choose the proposed model as a simple and sufficient approximation.

We add velocity parameters for each joint to the pose model to formulate the proposed motion model. Therefore, the human pose $P$ in a still frame $I$ is defined by the human scale parameter $S$ and the joint parameters $\left\{J^k\right\}_{k=1}^{K}$. The latter parameters include joint locations $Pos^k$ and velocities $V^k$ in the following form:

$$P = S \cup \left\{J^k\right\}_{k=1}^{K}$$

$$J^k = \left(Pos^k, V^k\right)$$

The objective function is similar to the one utilized by (Park and Ramanan, 2011):

$$Score(Pa) = \sum_t \Phi(P_t) + \sum_t \Psi(P_t, P_{t-1})$$

The temporal context is divided into two components:

$$\Psi(P_t, P_{t-1}) = \psi_s(S_{t-1}, S_t) + \sum_{k=1}^{K} \psi_j(J_t^j, J_{t-1}^j, S_{t-1})$$

The first component prefers the pose to have the same size in all frames of the input video sequence:

$$\psi_s(S_{t-1}, S_t) = -\frac{1}{2}\left(\frac{S_t - S_{t-1}}{S_{t-1}\sigma_s}\right)^2$$

And the second component has a form of a Linear Dynamical System:

$$\psi_j(P_t^j, P_{t-1}^j, S_{t-1}) = -\frac{dP_t^{j^T} \Sigma_p^{-1} dP_t^j}{2S_{t-1}^2}$$

$$dP_t^j = (P_t^j - AP_{t-1}^j)$$

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Sigma_p = \left[\begin{array}{c|c} \Sigma_p^p & \Theta \\ \hline \Theta & \Sigma_p^v \end{array}\right]$$

$$\Sigma_p^p = \alpha_p^{-1} I_{2\times2}$$

$$\Sigma_p^v = \alpha_v^{-1} I_{2\times2}$$

The proposed model corresponds to the constant motion model with presence of normally distributed error.

We want to notice that if the person scale parameter does not change in the video sequence ($S_1 = S_2 = \dots = S_N = S$), the basic model (Park and Ramanan, 2011) is a particular case of ours with the following parameters:

$$\alpha_p = 2\alpha S$$

$$\alpha_v \to \inf$$

## 2.4 Inference Algorithm

The proposed modification of the pose model makes dynamic programming unsutable for inference in such model. The main reason for this is a dependence of the velocity values on joint locations in all frames of the video sequence.

Therefore, we use a MCMC sampling technique to estimate the optimal value of the proposed score function. We sample a set of hypotheses from the distribution $p(Pa|W) \propto \exp(Score(Pa))$ to estimate the optimal set of poses in the input video sequence. We use the Metropolis-Hastings algorithm for sampling (alg. 1). The sample with the highest score is chosen as an approximation of the optimal solution. We want to notice that the algorithm requires a transition model $p(Pa'|Pa_{l-1})$ to sample from the specified distribution. It defines the way of construction new hypothesis $Pa'$ from the previous one $Pa_{l-1}$. An acceptance probability of the constructed hypothesis is computed in the following way:

$$Acc(Pa'|Pa_{l-1}) = \min\left(\frac{p(Pa'|W)p(Pa_{l-1}|Pa')}{p(Pa_{l-1}|W)p(Pa'|Pa_{l-1})}, 1\right)$$

It is important to notice that the Metropolis–Hastings sampling algorithm imposes insignificant limitations to a form of the optimized function. The only limitation is an existence of the partition function of the distribution $p(Pa|W)$. Therefore the proposed pose model in a video sequence can be extended with such global attributes of the observed human as a color of clothes and a physique.

### 2.4.1 Transition Model

The sampling algorithm requires the transition model $p(Pa'|Pa_{l-1})$ to sample from the specified distribution. All of the proposed steps of the transition model change only the pose joints locations and the scale parameters. Given the joint's locations we can optimally select hidden state values, as will be described in the next section.

In our experiments we use several types of steps to construct a new hypothesis of a set of human poses in the video sequence:

**Algorithm 1:** An approximate algorithm of the human pose estimation.

```
Data: $W, D$
Result: $Pa$
$Pa_0 = \{\underset{P_t}{argmax}\, p(P_t|I_t)\}_{t=1}^{N}$ ;
for $l$ = $1$ to $L$ do
    sample $Pa' \sim p(Pa'|Pa_{l-1})$;
    compute $Acc(Pa'|Pa_{l-1})$;
    sample $t \sim U(0,1)$;
    if $t < Acc(Pa'|Pa_{l-1})$ then
        $Pa_l = Pa'$;
    else
        $Pa_l = Pa_{l-1}$;
    end
end
$Pa = \underset{l \in \{1,2,...L\}}{argmax}\ p(Pa_l|W)$;
```

1. a random perturbation in human joint locations at the instant of time $t$;
2. a propagation of the human pose from the instant of time $t-1$ to the next instant of time;
3. a linear interpolation of the human pose in the interval $[t_1, t_2]$;
4. a replacement of the pose at the instance of time $t$ by one of hypothesis constructed by the algorithm of (Park and Ramanan, 2011).

All of the proposed steps require the time instance $t$ or the interval $[t_1, t_2]$ to be chosen. We want to notice that the choice of the interval is equivalent to choice of two instants of time. To speedup convergence we make the algorithm to prefer instances of time that have smaller confidence of the pose estimation correctness:

$$\xi(t) = \Phi(P_t) + \frac{1}{2}(\Psi(P_t, P_{t-1}) + \Psi(P_{t+1}, P_t))$$
$$p(t) \propto \max_{\tau} \xi(\tau) - \xi(t)$$

The first step type adds a small perturbations in the human joint locations and scale parameter at the instance of time $t$. This modification has the following form:

$$\begin{aligned} Pos_t^{j'} &= Pos_t^j + \beta S_t; \\ S_t{}' &= S_t + \gamma \\ \beta &\sim N(0, \beta_p^{-1} I_{2\times 2}) \\ \gamma &\sim N(0, \gamma_p^{-1}) \end{aligned} \qquad (2)$$

The second step type modifies a human pose at the instance of time $t$ by the propagation of its pose from the previous instance of time and addition a normally distributed noise according to (2). It uses the motion parameters to construct more likely hypothesis. For the first pose in the input video sequence this type of steps is equal to the previous one.

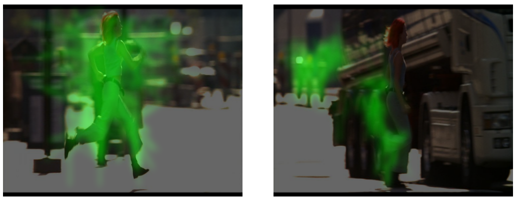

(a) (b)

Figure 2: Visualization of best constructed hypotheses. Area, where most hypotheses were found, are highlighted in green. In frame (a) detector find a lot of good hypotheses. In frame (b) detector cannot find a good set of hypotheses.

The third step type modifies a set of the human poses in the interval $[t_1, t_2]$. All poses inside the interval are constructed by a linear interpolation between the poses directly preceding and following the interval.

The fourth step type exploits only a set of hypotheses from the constructed by the human pose estimation algorithm in the still frame. It replaces a pose at instant of time $t$ with one of hypotheses from the constructed set. It prefers hypotheses that match the proposed temporal context, i.e. minimize $\Psi(P_t, P_{t-1})$ + $\Psi(P_{t+1}, P_t)$.

The fourth step type allows the inference algorithm to use high-scored poses found in a still frame. It speedups optimization in earlier stages. In figure 2 we demonstrate an area, where the best pose hypotheses were found. In other hand, it makes the inference sensitive to mistakes of human pose detector in a still frame (fig. 2 b). Therefore, we add the first three types of steps to deal with this problem.

### 2.4.2 Hidden State Estimation

As described above, the transition model modifies only joint locations and the scale parameters. Therefore the joint velocity values should be estimated. We choose the optimal values of the velocity parameters after each type of steps. In other words we chose values of the velocity parameters that maximize the score function:

$$V_l = \underset{V}{argmax}\, Score(Pa)$$

Velocities of the different joints are independent given the joint locations in the proposed model of the human pose in a video. Consequently, the velocity parameters of each joints can be estimated separately.

The form of the term $\psi_j(J_t^j, J_{t-1}^j, S_{t-1})$ implies that the function $Score(Pa)$ is factorized accordingly to the graphical model presented in the figure 3 given the human joint locations and the scale parameters. Therefore the velocity of each joint can be estimated efficiently.

We break the term $\phi_j(J_t^j, J_{t-1}^j, S_{t-1})$ into unary and pairwise terms based on the joint velocity parameters (we skips parameters of the terms to simplify description):

$$\psi_j = \psi_j^u(V_{t-1}^j) + \psi_j^p(V_{t-1}^j, V_t^j)$$

The unary and the pairwise terms have the following form:

$$\psi_j^u(V_{t-1}^j) = -\frac{{E^u}_t^{j^T} \Sigma_p^{p-1} {E^u}_t^j}{2S_{t-1}^2}$$

$${E^u}_t^j = \Delta Pos_{t-1}^j - A^u V_{t-1}^j$$

$$\Delta Pos_t^j = Pos_t^j - A_p^u Pos_{t-1}^j$$

$$\psi_j^p = -\frac{{E^p}_t^{j^T} \Sigma_p^{v-1} {E^p}_t^j}{2S_{t-1}^2}$$

$${E^p}_t^j = V_t^j - A^p V_{t-1}^j$$

$$A = \left[\begin{array}{c|c} A_p^u & A^u \\ \hline \Theta & A^p \end{array}\right] A_p^u, A^u, A^p \in \mathbb{R}^{2\times 2}$$

In this form the problem of a posterior velocity estimation is equal to the inference problem in the following Linear Dynamical System:

$$\overline{V}^j = \underset{V^j}{argmax}\, p\left(\overline{V}^j | \Delta\overline{Pos}^j\right)$$

$$\Delta\overline{Pos}_t^j \sim N(A^u \overline{V}_t^j, \Sigma_p^p)$$

$$\overline{V}_t^j \sim N(A^p \overline{V}_{t-1}^j, \Sigma_p^v)$$

$$\overline{V}_1^j \sim N(\mu_0, \Sigma_0)$$

$$\overline{V}_N^j = A^p \overline{V}_{N-1}^j$$

$$\mu_0 = \Theta_{2\times 1}$$

$$\Sigma_0 = \sigma_0 I_{2\times 2}$$

$$\sigma_0 \to \infty$$

where $\Delta\overline{Pos}^j$ is a set of observed normalized velocity values of the human joint, $\overline{V}^j$ is a set of normalized values of the hidden velocity parameters:

$$\Delta\overline{Pos}_t^j = \frac{\Delta Pos_t^j}{S_t}$$

$$\overline{V}_t^j = \frac{V_t^j}{S_t}$$

We use the Kalman filter with the RTS smoother (Rauch et al., 1965) to estimate the optimal values of the joint velocity.

We simulate a prior distribution on each component of the velocity parameters in the first frame with the normal distribution with dispersion going to infinity. Consequently, it indicates an absence of a prior preferences on the velocity. It implies the following modifications in the Kalman filter:

**Algorithm 2:** An algorithm of a joint velocity $V^j$ estimation.

```
Data: $\Delta\overline{Pos}_1^j, \ldots, \Delta\overline{Pos}_{N-1}^j$ are the observed
      data, $(A^u, A^p, \Sigma_p^p, \Sigma_p^v)$ are the model
      parameters
Result: $\mu_1^j, \ldots, \mu_N^j, \Sigma_1^j, \ldots, \Sigma_N^j$
// the Kalman filter;
$K_1 = A^{uT}\left(A^u A^{uT}\right)^{-1}$;
$\hat{\mu}_1^j = K_1 \Delta\overline{Pos}_1^j$;
$\hat{\Sigma}_1^j = \left(A^{uT} A^u\right)^{-1} A^{uT} \Sigma_p^p \left(A^u A^{uT}\right)^{-1} A^u$ ;
for $t$ = 2,..., N-1 do
    $\tilde{\Sigma}_{t-1} = A^p \hat{\Sigma}_{t-1}^j A^{pT} + \Sigma_p^v$;
    $K_t = \tilde{\Sigma}_{t-1} A^{uT} (A^u \tilde{\Sigma}_{t-1} A^{uT} + \Sigma_p^p)^{-1}$;
    $\hat{\mu}_t^j = A^p \hat{\mu}_{t-1}^j + K_t\left(\Delta\overline{Pos}_t^j - A^u A^p \hat{\mu}_{t-1}^j\right)$;
    $\hat{\Sigma}_t^j = (I - K_t A^u)\tilde{\Sigma}_{t-1}$ ;
end
// The RTS smoother;
$\mu_{N-1}^j = \hat{\mu}_{N-1}^j$;
$\Sigma_{N-1}^j = \Sigma_{N-1}^{\hat{j}}$;
for $t$ = N-2,...,1 do
    $K_t = \hat{\Sigma}_t^j A^p \tilde{\Sigma}_t^{-1}$;
    $\mu_t^j = \hat{\mu}_t^j + K_t(\mu_{t+1}^j - A^p \hat{\mu}_t^j)$;
    $\Sigma_t^j = \hat{\Sigma}_t^j + K_t(\Sigma_{t+1}^j - \tilde{\Sigma}_t^j) K_t^T$;
end
$\mu_N = A^p \mu_{N-1}^j$;
$\Sigma_N^j = \Sigma_p^v + A^p \Sigma_{N-1}^j A^{pT}$;
```

$$K_1 = A^{uT}\left(A^u A^{uT}\right)^{-1}$$

$$\hat{\mu}_1^j = K_1 \Delta\overline{Pos}_1^j$$

$$\hat{\Sigma}_1^j = \left(A^{uT} A^u\right)^{-1} A^{uT} \Sigma_p^p \left(A^u A^{uT}\right)^{-1} A^u$$

We estimate the values of $\overline{V}_t^j$ by the Viterbi algorithm (Viterbi, 1967) for LDS (alg. 2). It means that the original velocity parameters has the following estimations:

$$V_t^j | \Delta\overline{Pos}^j \sim N\left(S_t \mu_t^j, S_t^2 \Sigma_t^j\right)$$

## 3 RELATED WORKS

The mathematical model proposed in (Yang and Ramanan, 2011) is based on the deformable part model. We use it as a basic model of a human pose in a still frame.

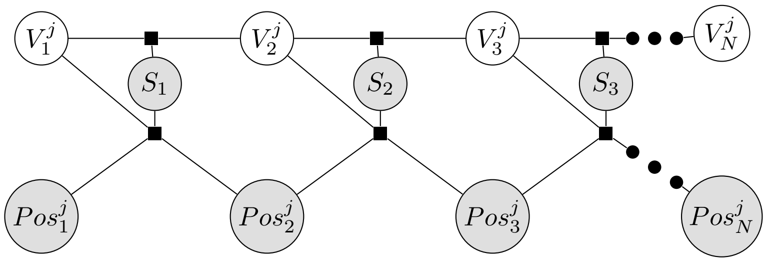


Figure 3: A graphical model for a joint velocity estimation. Nodes of the observed variables are shown in gray.

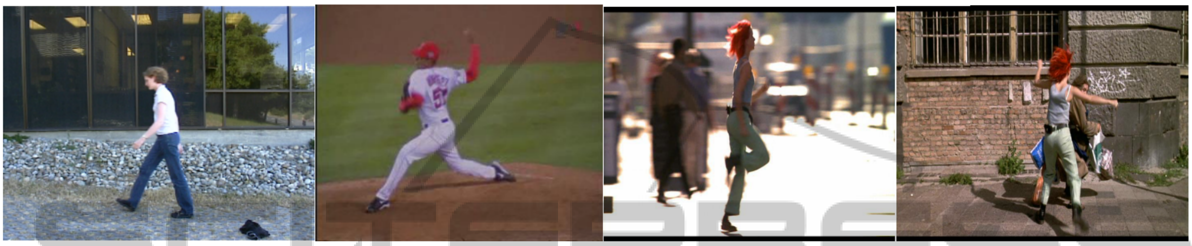


Figure 4: Frames from the dataset. From left to right, videos are called Walking, Pitching, Lola1, Lola2.

The deformable part model often fails in case of occlusions. It localizes each joint of the person based on evidence from the joint detector and location of the neighbours. In case of occlusions the detector is fully confident of the joint absence in its true position. One feasible solution of this problem was proposed in (Ghiasi et al., 2014). The authors extend number of detectors associated with each joint with detectors for occluded joints. The main disadvantage of this approach is a necessity of a prior knowledge of occlusion type. The authors consider only occlusions by another person. This approach can be applied in our framework in the future.

(Sapp et al., 2011) use the deformable part model for the human pose estimation in a video as well. In opposite to (Yang and Ramanan, 2011), they don't restrict the algorithm with sets of hypotheses constructed for each video frame separately. It makes the inference task much more complicated. In particular, they have to use an approximate inference algorithm.

Our algorithm of human pose estimation in a video isn't restricted to a choice from the set of hypotheses as well. The specific choice of the inference method was inspired by successful usage of the sampling technique for the tracking task (Shalnov and Konushin, 2013). In addition, it makes inference in complicated mathematical models possible. It allows further development of our model.

The deformable part model is not the only approach to human pose modeling. Recently the convolutional neural networks (CNN) has become widespread for image analysis. The authors of (Toshev and Szegedy, 2013) propose a CNN model of human pose in a still frame. The proposed algorithm achieved the best results in the standard datasets. Unfortunately, we cannot apply this approach in our model, because it doesn't allow neither construction of several hypotheses of a human pose, nor quality estimation of the outside specified pose.

(Girshick et al., 2014) proposes a way to construct an inference algorithm for the deformable part models as CNN. It makes possible to use an optimized and fast developing software tools (Jia et al., 2014) to human pose estimation task. In addition, it allows an adjustment of parameters for both the deformable part model and a feature extraction algorithm. We regard this approach as a most promising for the future development of the model.

# 4 EMPIRICAL EVALUATION

## 4.1 Setup

We quantitatively evaluate our algorithm on the publicly available dataset (Yang and Ramanan, 2011). Several frames from this dataset are shown in figure 4. The dataset includes 4 video sequences: **Pitching**, **Lola1**, **Lola2** and **Walking**. The videos are different in complexity. The video sequences **Pitching**, **Lola1** and **Lola2** contain motion of a camera. A zooming is presented in **Pitching**. **Lola2** contains several people in a scene.

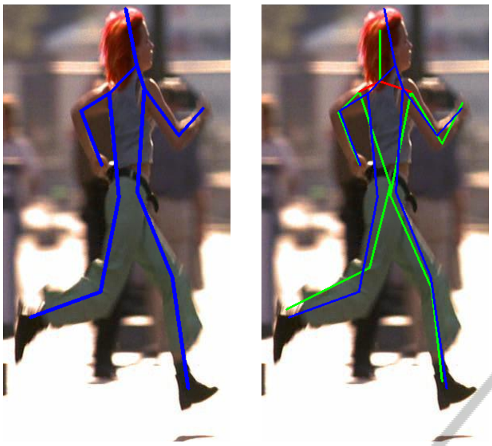


Figure 5: Pose as a set of sticks. (a) shows the groundtruth pose, (b) groundtruth and found poses. The ground truth pose is shown in blue. The green sticks are correctly localized accordingly to the PCP criterion, while the red sticks are considered as a false detections.

## 4.2 Results and Discussion

We evaluate the algorithm using the Percentage of Correct Parts (PCP) criterion introduced in (Ferrari et al., 2008). PCP criterion has a fundamental defect. It interprets a pose as a set of sticks (fig. 5 a) and classifies localization correctness for each of them independently. The stick is considered to be correctly localized if the normalized distance between found and groundtruth locations of its edges are within the specified threshold. The stick size is used as a normalization constant. Therefore, the same joint location can be correct for one stick and incorrect for another stick (fig. 5 b). However, PCP is the now-standard criterion for the human pose estimation task and we use it for evaluation.

For a fair evaluation we don't use tracking in the initialization stage. Therefore, the algorithm assumes that the prior distribution of human location is uniform in a video frame. We don't tune parameters for each video separately, on the contrary we use the same parameters for all videos. We presents results of comparison in figure 1. As a baseline we we use the method from (Park and Ramanan, 2011).

Table 1: We compare average PCP for the proposed method and the basic one. Results of the basic method are given from the (Park and Ramanan, 2011). Our approach outperforms the basic method in the most difficult video sequences.

| Algorithm | walking | pitching | lola1 | lola2 |
|---|---|---|---|---|
| basic | **0.950** | **0.797** | 0.670 | 0.500 |
| our | **0.950** | 0.762 | **0.695** | **0.545** |

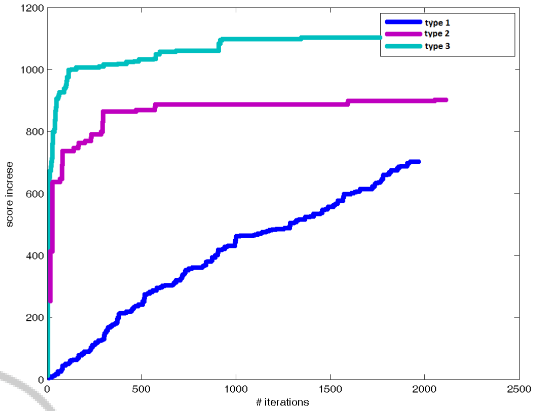


Figure 6: The diagram shows an impact of the different types of steps to the optimized score.

The proposed method outperforms the baseline in the most complicated scenarios **Lola1** and **Lola2**. The algorithm solves the ambiguity in presence of the several people in **Lola2**. It is achieved due to the information of the walking direction for the person of interest.

**Walking** is the simplest video in the dataset. The results of our algorithm and the basic one aren't differ in it. It is caused by limitations of the used human pose model in a still frame (Yang and Ramanan, 2011).

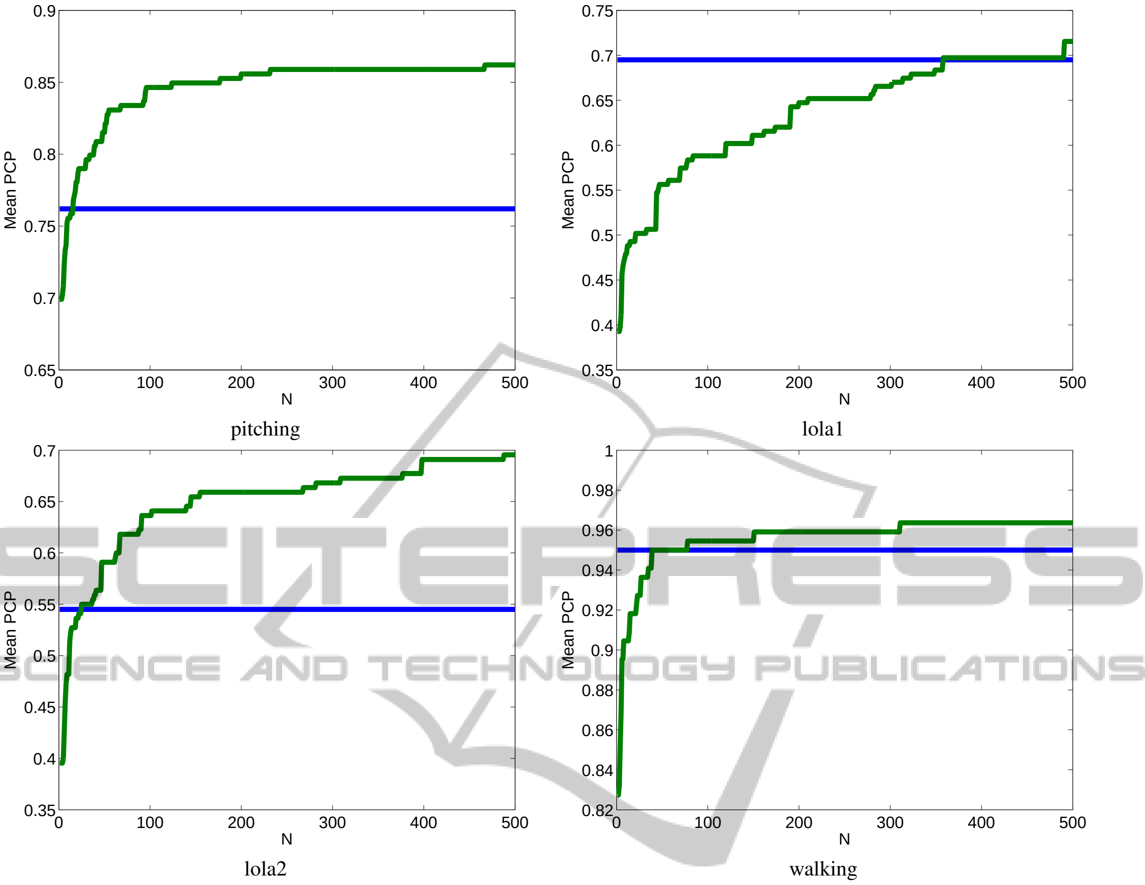
Our algorithm shows lower value of PCP criteria in the **Pitching** video sequence. The sophisticated motion of the human in the video sequence is the main reason of the algorithm failure.

In addtion, we evaluate the upper bound of human pose estimation in a video in case of using only hypotheses constructed by the detector of (Yang and Ramanan, 2011). The results are shown in figure 7. To evaluate the upper bound quality we construct a set of best hypotheses as described in (Park and Ramanan, 2011) and choose the best one in each frame using the PCP criterion.

The results show that our approach achieves almost optimal solution in **lola1** and **walking** videos. in **pitching** and **lola2** videos our approach cannot come near the optimal value. As we suppose the main reason of such behaviour is a precense of nonlinear motions of limbs. **walking** and **lola1** videos contain motion type that are more common in video surveillance.

## 4.3 Impact of the Different Steps

We evaluate an impact of different steps on the score increase during the inference (fig. 6). The results show that the second and the third types of steps produce the largest increase in the earlier stages of op-



pitching

lola1

lola2

walking

Figure 7: We show the upper bound on the mean PCP value of hypotheses (green curve) as a function of number of hypotheses ($N$). Our result values are shown as a blue line

timization. In the later stages the hypotheses constructed by those types of steps are rejected more frequently by the sampling algorithm (alg. 1). Therefore they bring the minor changes to the value of the optimized function in the latter stages of optimization.

The first type of step increases score equally throughout the optimization. But in each iteration of the inference algorithm the increase is relatively small.

## 5 CONCLUSIONS

In this paper we present the new mathematical model of a human pose in a video sequence. Our model is based on the model of human pose in a still frame proposed in (Yang and Ramanan, 2011). We expand this model with the new temporal context.

The proposed temporal context can be applied to any model of human pose in a still frame that allows:

- estimation of an arbitrary pose quality in a frame;
- construction of most likely pose hypotheses in a frame.

We show that the previous model (Park and Ramanan, 2011) is a particular case of the proposed one.

We propose the hidden state evaluation algorithm. The proposed algorithm estimates the values of hidden state parameters given joint locations and scale parameters.

We introduce a framework for a human pose estimation in a video based on the MCMC sampling technique. The proposed algorithm allows furthur development of our model with such global parameters of the person as a physique and a color of clothes.

Our model and inference algorithm produce better results in the most sophisticated videos of the dataset in comparison with the basic algorithm (Park and Ramanan, 2011).

## ACKNOWLEDGEMENTS

This publication is based on work funded by

Skolkovo Institute of Science and Technology (Skoltech), contract #081-R, Appendix A 2.

## REFERENCES

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Ghiasi, G., Yang, Y., Ramanan, D., and Fowlkes, C. C. (2014). Parsing occluded people. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2401–2408. IEEE.

Girshick, R., Iandola, F., Darrell, T., and Malik, J. (2014). Deformable Part Models are Convolutional Neural Networks.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM.

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140):269–294.

Park, D. and Ramanan, D. (2011). N-best maximal decoders for part models. *Computer Vision (ICCV), 2011 IEEE . . . .*

Rauch, H. E., Striebel, C., and Tung, F. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450.

Sapp, B., Weiss, D., and Taskar, B. (2011). Parsing human motion with stretchable models. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1281–1288. IEEE.

Shalnov, E. and Konushin, A. (2013). Improvement of mcmc-based video tracking algorithm. In *Pattern rcognition and image analysis (PRIA-11-2013)*, pages 727–730.

Toshev, A. and Szegedy, C. (2013). Deeppose: Human pose estimation via deep neural networks. *arXiv preprint arXiv:1312.4659*.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269.

Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts resenting shape. *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1385–1392.