Алгоритм для определения положения пользователей в мировых координатах и его применение для задач слежения[*]

А.Д. Грингауз, Е.В. Шальнов, А.С. Конушин

grin3s@mail.ru|shalnov.eugen@gmail.com|ktosh@graphics.cs.msu.ru

Факультет вычислительной математики и кибернетики

Московский государственный университет им.М.В.Ломоносова

В статье представлен многопользовательский алгоритм слежения (трекинга) , который работает в мировой системе координат. Этот алгоритм является модификацией подхода [1] и основан на минимизации энергии в течение временнóго скользящего окна. Он оценивает путь пользователя и его высоту в мировых координатах от известного места и положение его головы в координатах пространства изображения.

Ключевые слова: *алгоритм, трекинг, система слежения, мировые координаты, координаты изображения, 3D слежение, MCMC DA, минимизация энергии, оценка высоты.*

An algorithm for estimating locations of people in world coordinates and its application for tracking[*]

Alexander Gringauz, Eugene Shalnov, Anton Konushin

Lomonosov Moscow State University, Department of Computational Mathematics and Cybernetics

The paper presents a multi-person tracking algorithm that operates in world coordinates. This algorithm is a modification of the approach from [1] and is based on energy minimization within a temporal sliding window. It estimates a person's track and his height in world coordinates from known location of his head in image coordinates.

Keywords: *multi-person tracking algorithm, world coordinates, image coordinates, 3D tracking, MCMC DA, energy minimization, height estimation.*

Introduction

Multi-target tracking is an important computer vision task. It implies constructing trajectories for all people in a given video fragment. The trajectory contains a unique identifier for every person and his position in all frames of the video. This task is important for many applications, for example: video surveillance, improving pedestrian safety, marketing research. Despite a significant progress in recent years, humans are still far ahead of existing automatic algorithms in terms of solving this task.

All methods of object tracking can be divided two big groups: visual tracking methods and tracking by detection methods. Algorithms from the first group can be used for all kinds of objects and don't require an object detector. An object's position on the first frame must be set manually. An algorithm then finds objects on the following frames that are similar to the object on the first frame. The "Flock of features" algorithm [2] is an example of such methods.

Unlike the visual tracking methods the tracking by detection methods don't require an initialization of the object's position on the first frame. They use an object detector (for example, a person detector) to find objects and then associate detections belonging to a single object into a track. Greedy methods and bayesian models can be used to form tracks. The inference in bayesian models can be performed using MCMC [1] or graph cuts [3].

In case of a static camera with a known calibration matrix, we can estimate a person's 3D position in world coordinates. In [4, 5] the authors used multiple cameras. It allowed them to estimate 3D positions and height of all people more accurately. In [4] a 3D appearance model was introduced that allowed the algorithm to re-identify people more accurately, thus improving tracking performance. In [3] the authors showed that using 3D positions improved tracking performance. In [6] a particle-based tracker was used to construct trajectories on the ground plane in world coordinates.

The position and height of people is very important information that can be obtained from the scene. It can be used to determine the positions of people on a location map, which is useful for video surveillance systems.

We propose an approach to evaluate the location and height of the person in the world coordinates from the head location of the same person in different frames of the video. Then we embed this method into the person tracking algorithm [1]. Unlike other algorithms we do not use a full body detector to evaluate a person's location on the ground plane.

[*] Работа выполнена и опубликована при финансовой поддержке РФФИ, гранты 14-01-00849, 15-07-20347

Proposed Method

Brief description The proposed method is a modification of the approach from [1].

We choose the key frames from the whole set of frames $\{I_t\}$ with the interval *step*. In our experiments $step = 5$. Then a detector (see. section 44) is applied to every key frame to find heads. In the next step the algorithm constructs a tracklet (see. section 44) for each head detection. The tracklet is a combination of a head's location and its motion information in a small video segment. All tracklets from the current frame are added into the sliding window. Tracklets that went out of the sliding window are removed from it. After that the set of trajectories T is updated using new tracklets. They are added to existing tracks or form new ones. Then final results are formed for the frames that are between the key frame in the middle of the sliding window and the following key frame. A track contains a head's position on the image, legs' position and height of a person on every frame he was detected on. $T = \{T_i\}$, $T_i = \{(x_t^{(i)}, y_t^{(i)}, w_t^{(i)}, h_t^{(i)}, X_t^{(i)}, Y_t^{(i)}, Z_t^{(i)}, H_t^{(i)})\}$, $i = \overline{1, J}$, where $x_t^{(i)}, y_t^{(i)}, w_t^{(i)}, h_t^{(i)}$ - head's bounding box parameters on the frame t, $X_t^{(i)}, Y_t^{(i)}, Z_t^{(i)}$ - legs' position in world coordinates, $H_t^{(i)}$ - height.

In the following sections a more detailed description of the algorithm is provided.

Finding people

The proposed algorithm is aimed for people tracking. Therefore we use a HOG based head detector to detect people on each frame. It performs better than a full body detector in case of occlusions [7].

Building tracklets

The algorithm constructs a tracklet for each found head in the key frame. A tracklet is an object that contains information about the detection's location and its *velocity estimates* in a small video segment. We use the Flock of Features visual tracking algorithm [2] to construct it. It uses only one detection and tracks it forward and backward in time for several frames. Information about the detection includes a time stamp, parameters of the bounding box, its size and detector confidence. For more details see [1].

Estimation of a person's 3D position In this section we describe the modified part of the base algorithm. All other parts of the proposed method are almost the same as the corresponding parts of the base algorithm.

Our goal is to In order to embed three-dimensional coordinates into the algorithm. In order to achieve that we propose an algorithm to solve the task of estimating a single person's position and his height knowing his head's coordinates in the image.

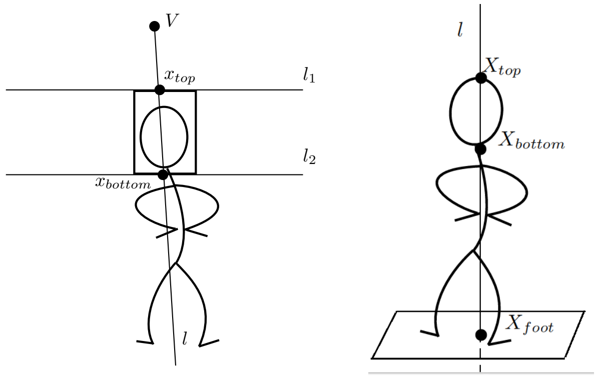

Figure 1: Visualization of the introduced variables.

Let $\{x_i, y_i, s_i\}, i = \overline{1, N}$ be the head's position and its size on several frames, N - number of frames the current person was detected on. P - the camera calibration matrix. A_{gp} - a perpendicular to the ground plane. R - the ratio of the person's height and his head's size (one of the algorithm's parameters, for all people it is the same). We use R equal to 6.5 in our experiments. Formally, our goal is to estimate the three-dimensional position of the person on A_{gp}: $\{X_i, Y_i, Z_i\}, i = \overline{1, N}$ and his height H.

Let's consider two approaches to solve this task: independently on each frame, jointly on all frames.

Position estimation on a single frame Lets introduce notations:

- $X_{foot} = [X, Y, Z, W]^T$ are the homogeneous coordinates of the person's feet in the world coordinates.
- X_{top} are the homogeneous coordinates of the top point of the person's head in the world coordinates.
- x_{top} are the homogeneous coordinates of the top point of the person's head on the image.
- H is the person's height in world coordinates.
- X_{bottom} are the homogeneous coordinates of the bottom point of the person's head in the world coordinates.
- x_{bottom} are the homogeneous coordinates of the bottom point of the person's head on the image.
- l_1 is the horizontal line on the image that goes through x_{top}.
- l_2 is the horizontal line on the image that goes through x_{bottom}.
- l is the projection of the vertical line going through X_{foot} onto the image.

These notations are shown on the figure 1.

The lines l_1 and l_2 can be computed by selecting two points on the head's bounding box. To compute the line l we must find the vanishing point of the vertical lines that are orthogonal to the ground plane. It can be found as shown in [8]:

$$V = PA_{gp} \quad (1)$$

Thus, l can be found using V and the center of the head's bounding box.

X_{foot} lies in A_{gp}:

$$A_{gp}X_{foot} = 0 \tag{2}$$

Normalization constraint on the last coordinate of X_{foot}:

$$W = 1 \tag{3}$$

The top point of the head can be found as:

$$X_{top} = X_{foot} + HA_{gp} \tag{4}$$

x_{top} is a projection of X_{top} onto the image:

$$x_{top} = PX_{top} \tag{5}$$

The bottom point of the head can be found as:

$$X_{bottom} = X_{foot} + \frac{R-1}{R}HA_{gp} \tag{6}$$

x_{bottom} is a projection of X_{bottom} onto the image:

$$x_{bottom} = PX_{bottom} \tag{7}$$

x_{top} lies on l_1:

$$l_1 x_{top} = 0 \tag{8}$$

x_{bottom} lies on l_2:

$$l_2 x_{bottom} = 0 \tag{9}$$

x_{top} lies on l:

$$l x_{top} = 0 \tag{10}$$

Let's solve the system (2) - (10) with respect to X, Y, Z, H. It can be transformed into the system of linear equations with the square non-singular matrix M and the right side b:

$$MX_{res} = b,\ X_{res} = [X, Y, Z, H]^T\ X_{res} = M^{-1}b \tag{11}$$

Estimating position on multiple frames If the person's positions on multiple frames are known then we can use the algorithm described in the previous section for every frame separately and average the values of height, computed for each frame.

We consider the person's height to be a constant for each person (if his doesn't jump, crawl or lie down). This assumption can be used to smooth errors that come from solving the system (11) for each frame separately. Despite the fact that the system (11) can be solved exactly, errors occur due to slightly inaccurate results produced by the detector.

This problem can be formulated as the system of linear equations that is a union of systems of the form (11) for every frame with a common variable H, corresponding to the person's height.

$$M_{all}X_{all} = b_{all},\ X_{all} = [X_1, Y_1, Z_1, \cdots, X_N, Y_N, Z_N, H]^T \tag{12}$$

The size of M_{all} is $4N \times (3N+1)$, therefore the system (12) is over-determined and can only be solved approximately. One of possible approaches to solve it is to minimize the norm of the residual $\|M_{all}X - b_{all}\|$. In the current work we used the L-1 norm because it is more stable than the L-2 norm if there are outliers in data. We reduced this optimizations problem to the linear programming problem and solved it with the simplex method.

Our experiments show that in order to achieve the best results it is necessary to set different weights for the equations (2) - (10), that is to multiply them by a certain number. These weights show how strong the impact of every equation is on the residual while approximately solving (12). Those equations that must be satisfied very precisely should have larger weights compared to the rest. The equations (2) and (10) were multiplied by 1000, others were left without changes.

Building trajectories In this section an algorithm to combine tracklets into trajectories is described.

Let D be a set of all detection in the sliding window. H - a hypothesis that defines how D is divided into trajectories, that is: $H = \{T_1, ..., T_J\}$, $T_i = \{d_n^j\}$, d_n^j - the n^{th} detection of the j^{th} trajectory. Now we define a generative model for detections and trajectories:

$$p(D, H) = p(D|H)p(H) \tag{13}$$

We need to find:

$$H^* = argmax_H p(D, H) \tag{14}$$

This probabilistic model is almost the same is the model from [1]. The inference is performed using MCMC. The only difference is that the legs' position in world coordinates is used while computing the likelihood of the trajectory hypothesis. The legs' positions are computed using the algorithm from the section 44. Besides, it is necessary to translate the velocity estimates in tracklets from the image coordinates to the world coordinates. For more details see [1].

Experimental evaluation

The aim of the experimental evaluation is to compare the base algorithm [1] with its modification. The modified part consists in estimating people's positions and height in world coordinates and using this information for tracking (see sections 44 and 44). We want to find out how the usage of world coordinates affects tracking performance.

We use the TownCenter dataset for numerical comparison of the proposed algorithm with the basic one. It is a high resolution (1920 × 1080/25fps) video, framed from a static camera. The calibration matrix for the camera and the ground truth are provided. The method is implemented in MATLAB using C++ MEX-functions. The time the algorithm needs to process 1000 frames using the head detector once in 5 frames is ∼ 4 hours on a computer with Inter Core i7, 12 GB RAM.

To evaluate the quality of the algorithms standard metrics like precision and recall along with the CLEAR MOT [10] group of metrics were used: FP - number of false positives, FN - number of false negatives, ID - number of identity switches, MOTA - a total error that takes into account FP, FN and ID; MOTP shows how close the trajectory lies to the real person's position obtained from the ground truth.

The results are shown in the table 1. All metrics were computed in the image space.

Algorithm	Base	Modification
Precision	80.7	71.4
Recall	74.7	75.6
FP	2609	4407
FN	3696	3567
ID	22	43
MOTA	56.6	45.0
MOTP	47.2	47.2

Table 1: Comparison of the base and the modified algorithm.

Analyzing results

The experimental results show that the proposed method can successfully estimate a person's leg position and his height in world coordinates. However, the proposed method shows slightly lower accuracy than the base method. The main drawback of the proposed method of finding positions of people in world coordinates is the necessity to know the ratio between a person's height and his head's size. The head detector produces bounding boxes that may have different size. They may be shifted from the head's actual position. These reasons can lead to errors in estimating the leg position and height.

Conclusion

In this work we proposed a new method of estimating the position in world coordinates and height of a person using his head's position on the image. Based on this method we design a modified version of the people tracking algorithm that can find trajectories and persons' heights in world coordinates. Despite some shortcomings, this approach allows us to use 3D positions of people for tracking. The further development of the algorithm may include: code optimization, adding new factors to the probabilistic model (appearance, foreground [12]), expanding the method in order to process scenes filmed from multiple cameras [11] (in this case we don't need to know the ratio between the head size and the full body size).

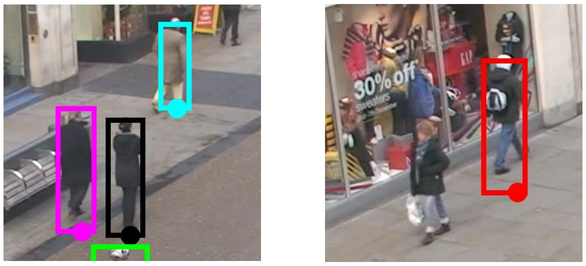
Figure 2: An example of tracking results.

Bibliography

[1] Shalnov E., Konushin A. Improvement of MCMC-based video tracking algorithm. Pattern recognition and image analysis (PRIA-11-2013), 2013, p.727-730.

[2] Kolsch M., Turk M. Fast 2d hand tracking with flocks of features and multi-cue integration //Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on. – IEEE, 2004. – p. 158-158.

[3] Andriyenko A., Schindler K., Roth S. Discrete-continuous optimization for multi-target tracking //Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. – IEEE, 2012. – p. 1926-1933.

[4] Baltieri D. et al. Multi-view people surveillance using 3D information //Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. – IEEE, 2011. – p. 1817-1824.

[5] Berclaz J. et al. Evaluation of probabilistic occupancy map people detection for surveillance systems //Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. – 2009. – num. LIDIAP-CONF-2009-064.

[6] Batanov P., Kononov V., Konushin A. People tracking in surveillance systems for sport games using multiple cameras. Graphicon, 2013, pp. 333-336 (in Russian).

[7] Konushin A., Filippov I., Konushin V., Kononov V. Counting people in a video sequence based on head detection. Programmnye produkty i sistemy, no. 1, 2015, pp. 121-126 (in Russian).

[8] Hartley R., Zisserman A. Multiple view geometry in computer vision. – Cambridge university press, 2003.

[9] Milan A., Roth S., Schindler K. Continuous energy minimization for multitarget tracking //Pattern Analysis and Machine Intelligence, IEEE Transactions on. – 2014. – Vol. 36. – num. 1. – p. 58-72.

[10] Keni B., Rainer S. Evaluating multiple object tracking performance: the CLEAR MOT metrics

//EURASIP Journal on Image and Video Processing. – 2008.

[11] Kononov V., Konushin V., Konushin A. People Tracking Algorithm for Human Height Mounted Cameras. DAGM, 2011, pp. 163-72.

[12] Chetverikov N., Konushin A. Finding objects in a video stream using graph cuts. Graphicon, 2012, pp. 262-265 (in Russian).

[13] Gringauz A., Shalnov E., Konushin A. Modification of the Multi-target Tracking Algorithm Based on Energy Minimization. GraphiCon-2014, 2014, pp.139-142.