Восстановление позы и комплекции человека по видео

Шалимова Екатерина Алексеевна

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: ekaterina.shalimova@graphics.cs.msu.ru

Научный руководитель *— Шальнов Евгений Вадимович*

Одним из ключевых препятствий на пути развития методов распознавания поведения человека по видео является ограниченность эталонных данных. Большинство существующих эталонных коллекций получено съёмкой только с одного ракурса. Использование методов захвата движения по видео, восставливающих позу и фигуру человека на каждом кадре, позволит синтезировать новые ракурсы и расширить существующие датасеты. Однако существующие методы обладают недостаточной точностью и не обеспечивают временну́ю согласованность (гладкость) движения. В данной работе рассматривается задача восстановления позы и комлекции человека по видео на основе оценки параметров трёхмерной модели человека SMPL [1] с обеспечением временно́й согласованности параметров.

Наиболее точный на данный момент метод восстановления позы и комплекции человека (SPIN [2]) предназначен для применения к одному изображению. При покадровом применении алгоритма SPIN к видеопоследовательности на соседних кадрах наблюдаются нереалистичные изменения положения отдельных суставов и поворота фигуры в целом. Кроме того, при быстром изменении положения частей тела человека отдельные кадры видеопоследовательности получаются смазанными, из-за чего базовый метод допускает на таких кадрах заметные ошибки.

Для устранения вышеперечисленных недостатков базового метода в работе предлагается оптимизировать его покадровые результаты с использованием информации о положении фигуры человека на предшествующих и последующих кадрах видеопоследовательности. Предложено минимизировать с помощью градиентного спуска следующую функцию потерь:

$$L = \sum_{i=2}^{N_f} (\hat{X}_i - \hat{X}_{i-1})^2 + \sum_{i=1}^{N_f} (\hat{x}_i - x_i)^2 + L_{prior}, \qquad (1)$$

$$L_{prior} = -\sum_{i=2}^{N_f} \log(\sum_{k}^{N_k} w_k \mathcal{N}(\hat{X}_i - \hat{X}_{i-1}; \mu_k, \sigma_k)), \tag{2}$$

где N_f — количество кадров видеопоследовательности, $\hat{X}_i$ — трёхмерные координаты ключевых точек модели человека на кадре i, $\hat{x}_i$ — проекция трёхмерных координат ключевых точек фигуры на кадре i на плоскость изображения, x_i — двумерные координаты ключевых точек фигуры на кадре i, предсказанные сторонним детектором позы. Для оценки правдоподобия изменения трёхмерных координат между соседними кадрами на наборе данных CMU MoCap [3] была обучена смесь нормальных распределений с числом компонент $N_k = 8$. L_{prior} (2) представляет собой оценку правдоподобия изменения трёхмерных координат между соседними кадрами для данной смеси нормальных распределений, где w_k, μ_k и σ_k — априорная вероятность и параметры нормального распределения компоненты смеси k.

Наиболее информативной метрикой качества результата могло бы быть расстояние между предсказанными и истинными трёхмерными координатами ключевых точек фигуры человека. Однако поскольку в имеющихся наборах данных походок нет информации о трёхмерных координатах ключевых точек фигуры, предлагается оценивать результат с помощью двух следующих метрик:

$$RE = \frac{1}{N_f} \frac{1}{N_j} \sum_{i=1}^{N_f} (\hat{x}_i - x_i)^2, \tag{3}$$

$$TSm = \frac{1}{N_f - 1} \sum_{i=2}^{N_f} (\hat{X}_i - \hat{X}_{i-1})^2, \tag{4}$$

где $\frac{1}{N_j} = 19$ — количество ключевых точек фигуры человека. Метрика RE (ошибка реконструкции) представляет собой сумму расстояний между координатами проекций ключевых точек фигуры на плоскость изображения $\hat{x}_i$ и двумерными координатами ключевых точек фигуры x_i, предсказанными детектором позы. Уменьшение ошибки реконструкции соответствует улучшению соответствия между предсказаниями детектора позы и метода восстановления позы и комплекции человека. Метрика TSm (временна́я гладкость предсказаний) представляет собой сумму расстояний между трёхмерными

координатами ключевых точек фигуры на соседних кадрах. Уменьшение TSm соответствует уменьшению расстояния между ключевыми точками фигуры на соседних кадрах и устранению неправдоподобных скачков положения конечностей, вызванных ошибками базового метода.

Оценка работы предложенного алгоритма производилась на наборе данных TUM GAID [4]. В таблице 1 приводятся значения метрики для исходных предсказаний и для предсказаний, оптимизированных с помощью алгоритма.

Таблица 1: Средние значения метрик RE и TSm на наборе данных [4] до и после оптимизации

	RE, пикселей	TSm, см
Значение до оптимизации	10.75	81.89
Значение после оптимизации	**8.22**	**67.92**

Литература

1. Loper M., Mahmood N., Romero J., Pons-Moll G., Black M. SMPL: A Skinned Multi-Person Linear Model // In ACM TOG (Transactions on Graphics), 2015, vol. 34, no. 6, pp. 1-16.
2. Kolotouros N., Pavlakos G., Black M.J., Daniilidis K. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop // In Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2252-2261.
3. Набор данных CMU MoCap: `http://mocap.cs.cmu.edu`
4. Hofmann M., Geiger J., Bachmann S., Schuller B., Rigoll G. The TUM Gait from Audio, Image and Depth (GAID) Database: Multimodal Recognition of Subjects and Traits // In Journal of Visual Communication and Image Representation, Special Issue on Visual Understanding and Applications with RGB-D Cameras, vol. 25, no. 1, pp. 195-206, Elsevier, 2014.