Вопросно-ответная система с автогенерацией пространственных структур

Воронов Сергей
Научный руководитель: Б.Г. Кац, К.В. Воронцов

Московский Физико-Технический Институт
Факультет Управления и Прикладной Математики
Кафедра Интеллектуальных Систем

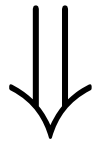
Москва
2016

Постановка задачи

Глобальная задача: научиться отвечать на всевозможные вопросы об объектах на изображениях и видео неотличимо от человека ("Тест Тьюринга для компьютерного зрения")

Постановка задачи

Глобальная задача: научиться отвечать на всевозможные вопросы об объектах на изображениях и видео неотличимо от человека ("Тест Тьюринга для компьютерного зрения")



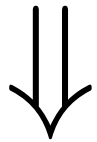
Задача: научиться генерировать траектории объектов, чьи движения описаны в тестовой форме, и отвечать на вопросы о них

Обзор литературы

- Идея одновременного поиска треков-состояний моделей и общая архитектура
  Yu H. et al.«A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video.»
- Снятие многозначности предложений при помощи поясняющих видео (например, «Человек подошел к сумке с рюкзаком.»)
  Berzak Y. et al. «Do You See What I Mean? Visual Resolution of Linguistic Ambiguities.»
- Предсказание значения слова по изображениям, с целью дальнейшего использования для описания изображения
  Mao J. et al. «Learning like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images.»

Подзадачи

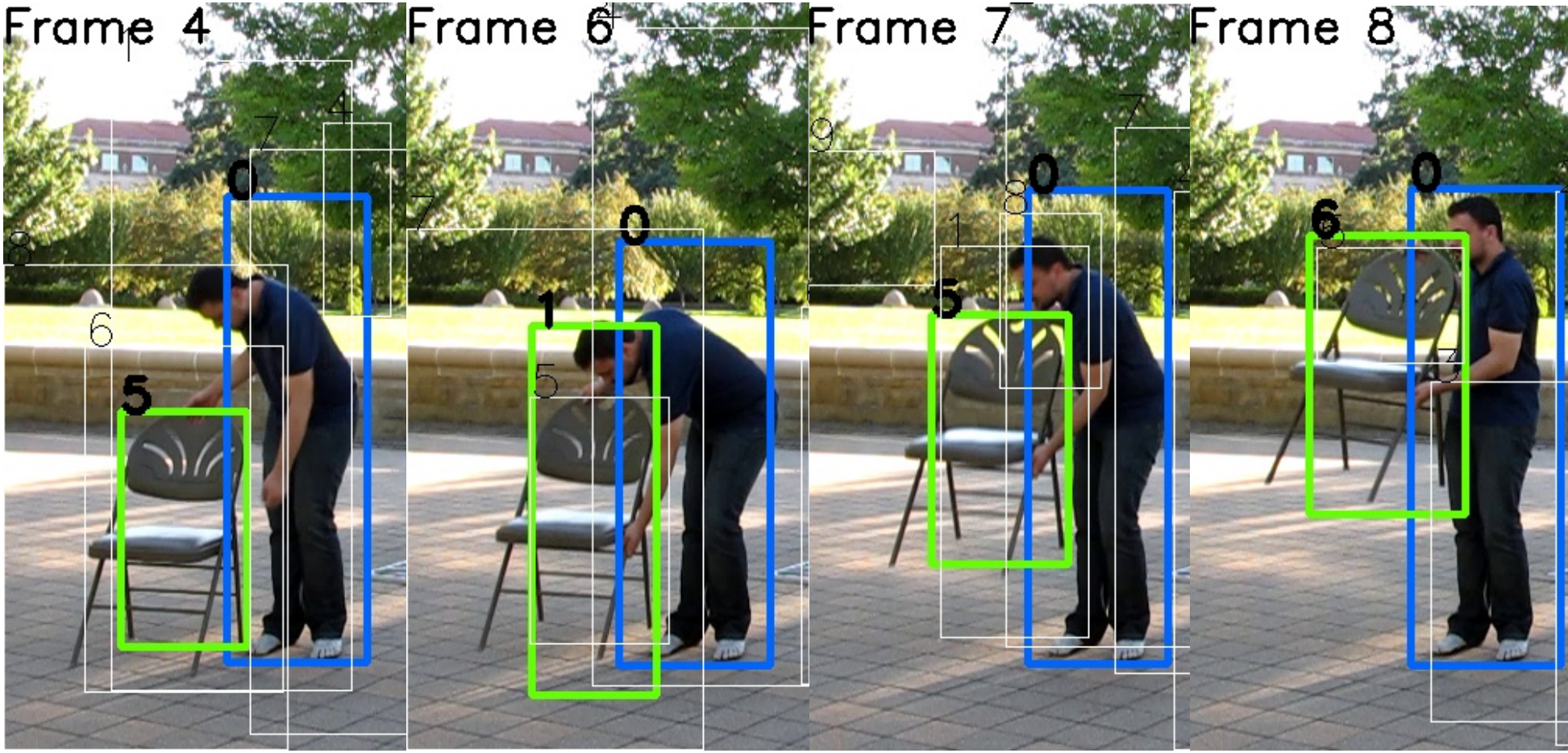
Задача: научиться генерировать траектории объектов, чьи движения описаны в тестовой форме, и отвечать на вопросы о них



- Построить функцию качества для пары (предложение, видео)
- Научиться генерировать с помощью данной функции траектории объектов из текста

Траектория (трек)

Трек – набор найденных позиций объекта на кадрах (по одной на кадр)



Обнаружение треков

j – трек объекта
b_j^t – позиция объекта, выбранная для трека j на кадре t
$f(b)$ – качество обнаружения объекта b
$g(b, b') = -[\text{dist}(b, b') - \text{optical\_flow}(b, b')]$ – расстояние между объектами на соседних кадрах

$$j = \arg\max_j \left(\sum_{t=1}^{T} f(b_j^t) + \sum_{t=2}^{T} g(b_j^{t-1}, b_j^t) \right)$$

Признаки

Используются для получения специфичной информации о перемещении объектов.

$$f : \{ObjectPositions\} \to \{1..k\}$$

Примеры:

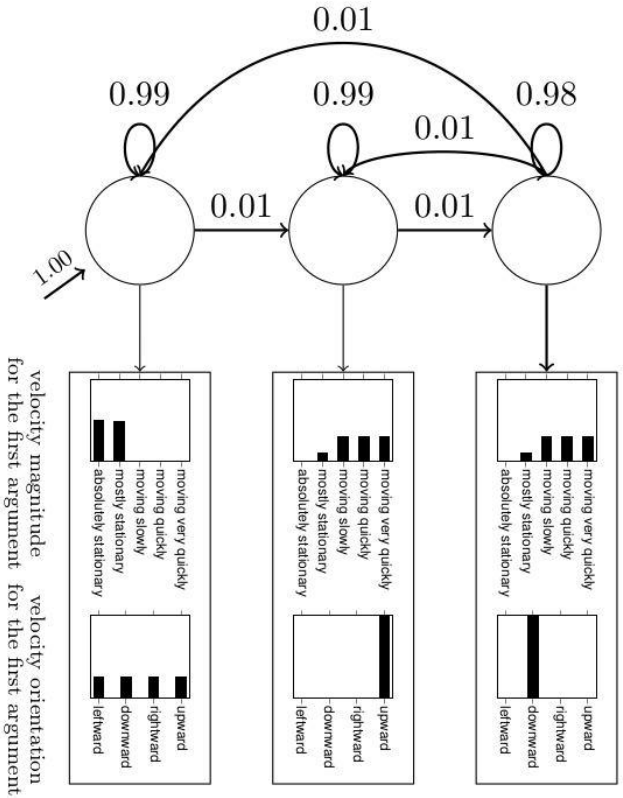
- Скорость
- Относительное расстояние
- Направление скорости

Зависит от:

- одного или двух объектов
- их позиций на кадрах $t-1, t$

Представление слова

- многие действия можно разбить на составные части
- закодируем части действия как одиночные состояния HMM
- каждое состояние **Словарной Модели** содержит распределения вероятностей значений признаков



Изображение из Yu et al. A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video.

Скоринг: (треки, **Словарная Модель**)

Пусть m – **Словарная Модель** с двумя объектами.
Правдоподобие пары (кадр fr, состояние k):

$$L_{st-fr}(k, fr) = \prod_{feat \in k} feat(obj_1, obj_2) \prod_{feat_1 \in k} feat_1(obj_1) \prod_{feat_2 \in k} feat_2(obj_2)$$

Скоринг: (треки, **Словарная Модель**)

Пусть m – **Словарная Модель** с двумя объектами.
Правдоподобие пары (кадр fr, состояние k):

$$L_{st-fr}(k, fr) = \prod_{feat \in k} feat(obj_1, obj_2) \prod_{feat_1 \in k} feat_1(obj_1) \prod_{feat_2 \in k} feat_2(obj_2)$$

$tr = t \to k^t$, где k^t – состояние **Словарной Модели** m на кадре t

$$L(m, \text{frames set}, tr) = \prod_{t=1}^{n} L_{st-fr}(k^t, fr_t) \prod_{t=2}^{n} hmm(k^{t-1}, k^t)$$

$$\log(L) = \sum_{t=1}^{n} h(k^t, fr_t) + \sum_{t=1}^{n} a(k^{t-1}, k^t),$$

где $h(k^t, fr_t)$ показывает насколько хорошо состояние модели описвает кадр и $a(k^{t-1}, k^t)$ – вероятности перехода между состояниями модели

Скоринг: (позиции объектов, **Словарная Модель**)

Пусть у нас теперь L треков и M моделей.

$$LL = \max_{\mathbf{J},\,\mathbf{K}} \left[\sum_{l=1}^{L} \left(\sum_{t=1}^{T} f(b_{j_l}^t) + \sum_{t=2}^{T} g(b_{j_l}^{t-1}, b_{j_l}^{t-1}) \right) \right.$$

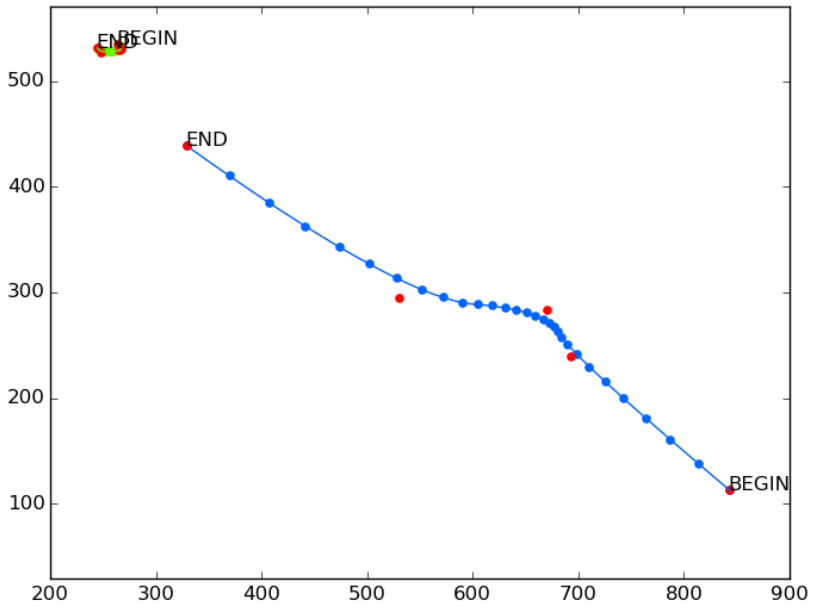
$$\left. + \sum_{m=1}^{M} \left(\sum_{t=1}^{n} h(k_m^t, fr_t) + \sum_{t=1}^{n} a(k_m^{t-1}, k_m^t) \right) \right]$$

Генерация треков

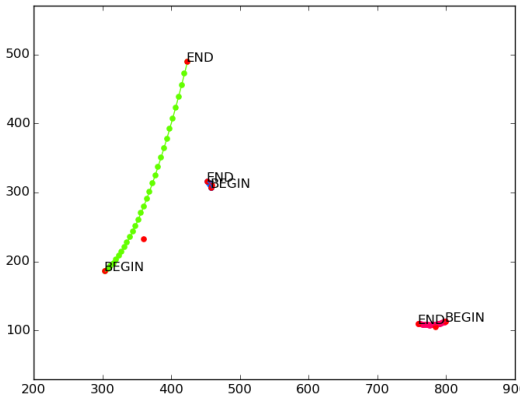
Трек содержит одну позицию объекта на каждом кадре

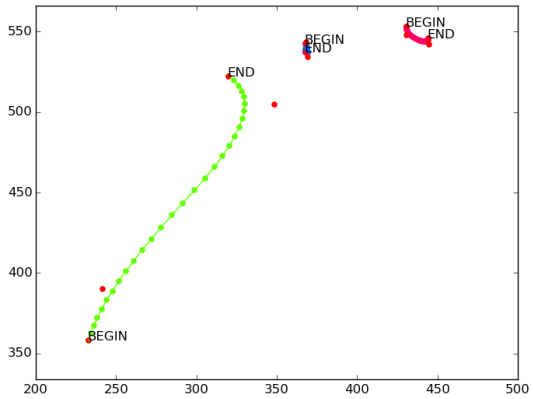
«A person approached a stationary bicycle»

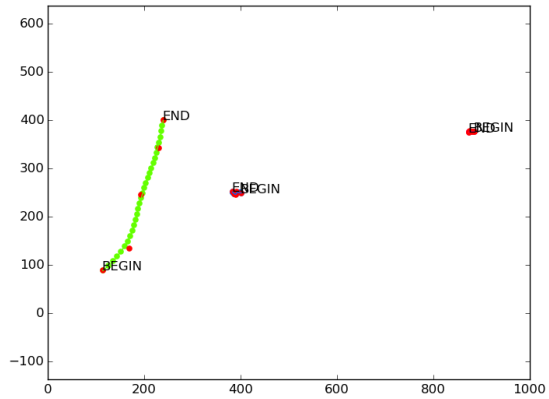
- Каждый трек – это B-сплайн с небольшим количеством контрольных точек
- Начинаем с набора случайных треков и пересемплируем контрольные точки
- Функция качества – это правдоподобие с $g = 0$



Зависимость кривизны треков от разного количества котрольных точек



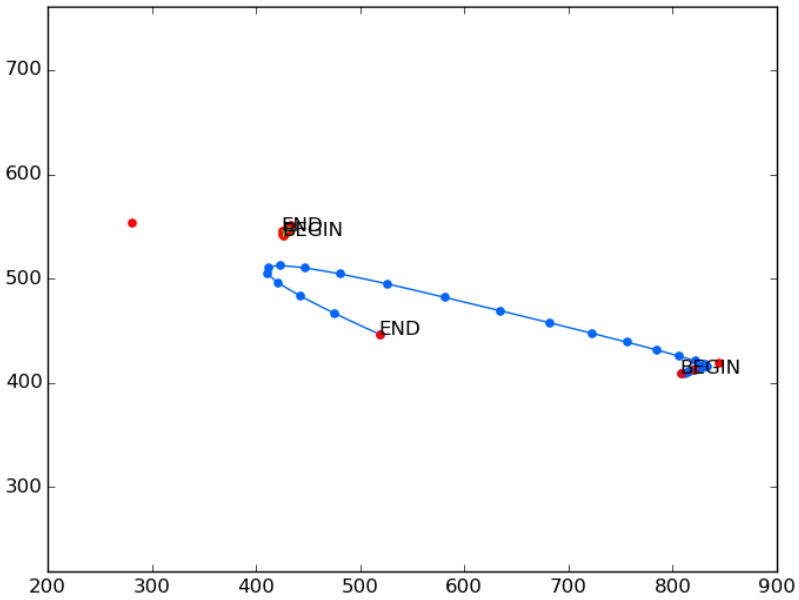




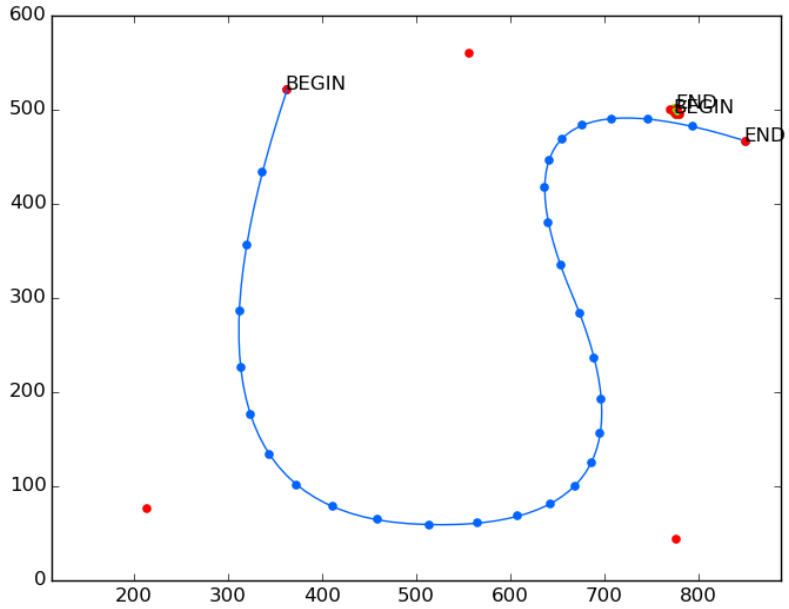
Неподвижный человек поднял кресло, находясь слева от неподвижного велосипеда.

A stationary person picked up a chair to the left of stationary bicycle.

Треки для похожих предложений



(a) A **slow** person came up to a stationary chair



(b) A **fast** person came up to a stationary chair

Разница в треках для предложений с одним измененным словом.

Общий подход

- Получить информацию об объектах, используемых в предложении и вопросе, используя парсер START
- Сгенерировать 3 видеоподобных объекта
- Используя абсолютные положения объектов, найти ответ на вопрос
- Сгенерировать возможные ответы, используя шаблоны вопросов

Типы вопросов

	2-об.	3-об.
How far is obj_1 from obj_2 in time t?	70%	49%
Where was obj in time t?	64%	41%
Is obj_1 left of the obj_2 in time t?	88%	60%
Is obj_1 right of the obj_2 in time t?	88%	60%

Заключение

- Предложен алгоритм генерации треков для движений объектов, описанной в текстовой форме
- Предложено использование этого алгорима для ответа на вопросы о прострaственной структуре объектов
- Показана работоспособность данного алгоритма