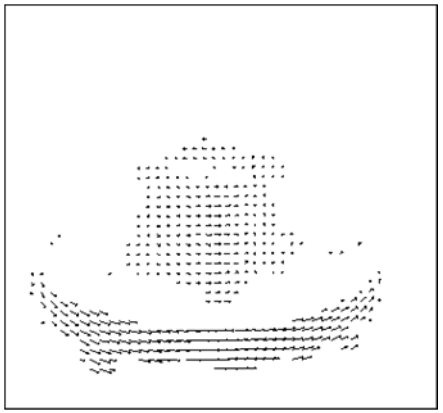
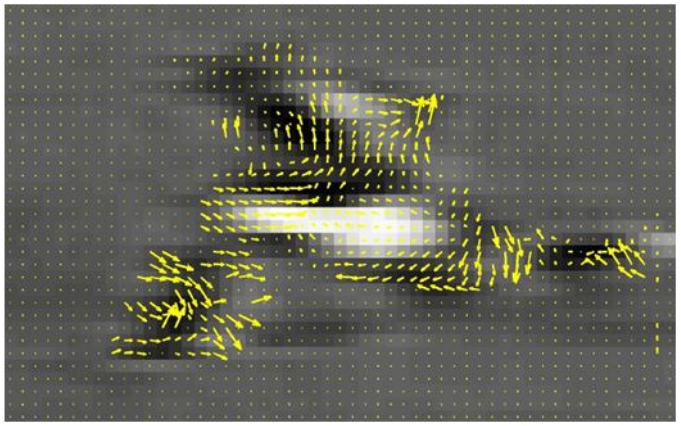
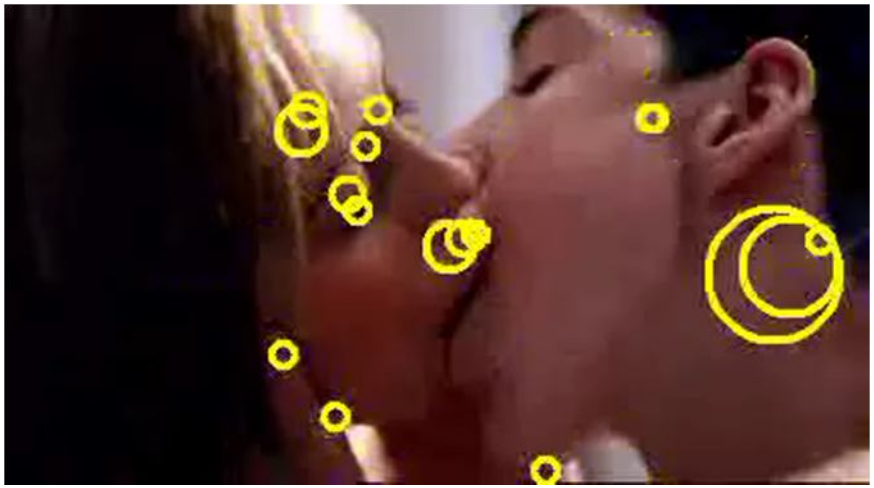
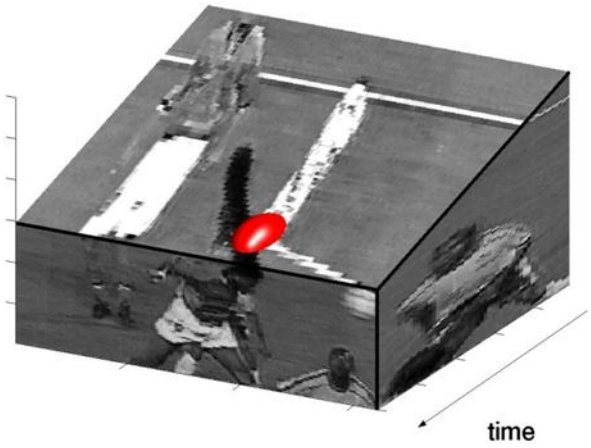
Распознавание действий в видео









Many slides adapted from Rick Szeliski, Alexei Efros and Ivan Laptev

Общая информация

Microsoft Research

Этот курс подготовлен и читается при поддержке

Microsoft Research

- Страница курса

http://courses.graphicon.ru/main/vision

Распознавание действий

- Огромное количество видеороликов
- Действия (action) людей – главные события в кино, новостях, домашнем видео, видеонаблюдении

BBC Motion Gallery

YouTube Broadcast Yourself

150,000 uploads every day





Для чего может пригодится:

- Навигация по контенту
  - *Перемотка до следующей важной сцены (пр. гола)*
- Поиск видео
  - Найти сцену «Обама пожимает руку Медведеву»
- Социальные науки
  - *Влияние сцен курения в кино на подрастающее поколение*

Действия человека

Определение 1:

- Физическое движение тела человека



KTH action dataset

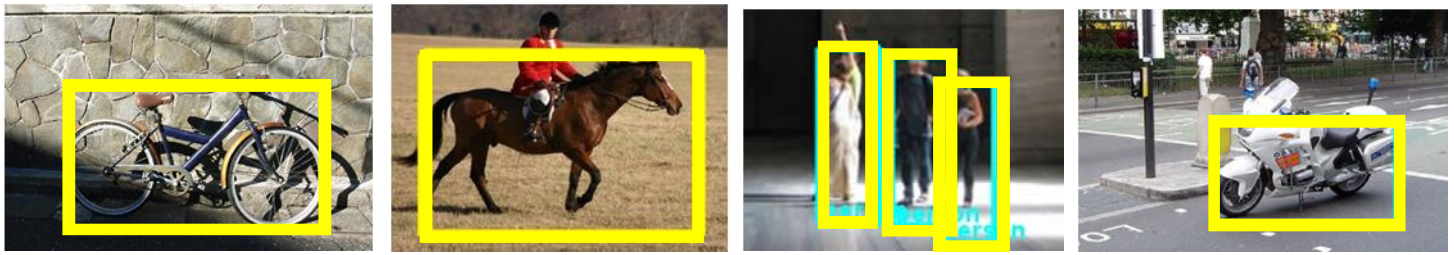
Определение 2

- Взаимодействие с окружением с определенной целью
  - *Одно и то же движение имеет разный смысл в зависимости от окружения*




Распознавание действий

“стабильные по виду” объекты

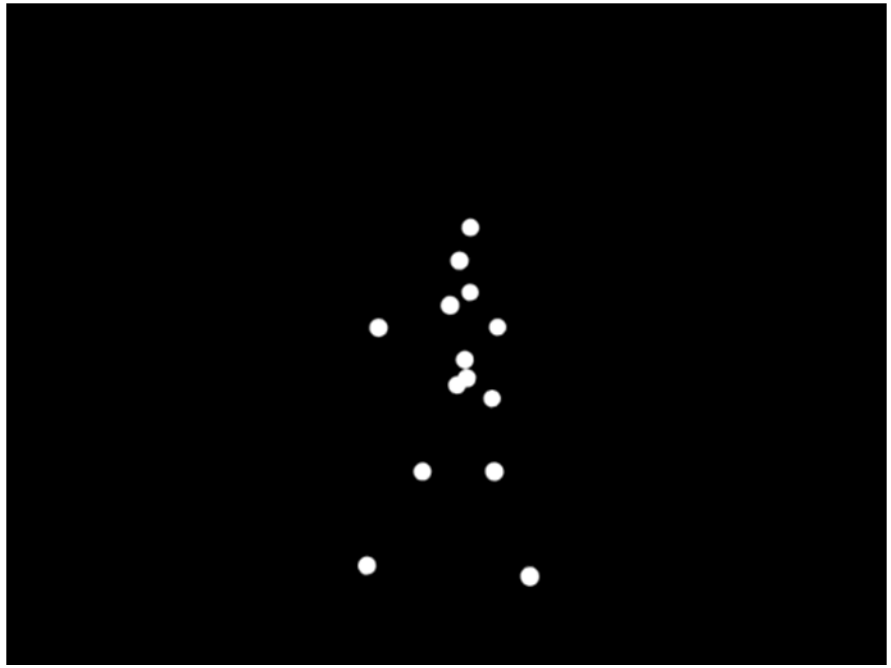


“атомарные действия

car exit phoning smoking hand shaking drinking

- Человек может некоторые действия распознать по статичному изображению
- Много сложностей:
  - Все обычные сложности при распознавании
  - Некоторые действия не возможно определить по одному кадру
  - Для описания действия нужно распознавать несколько предметов и их взаимное отношение

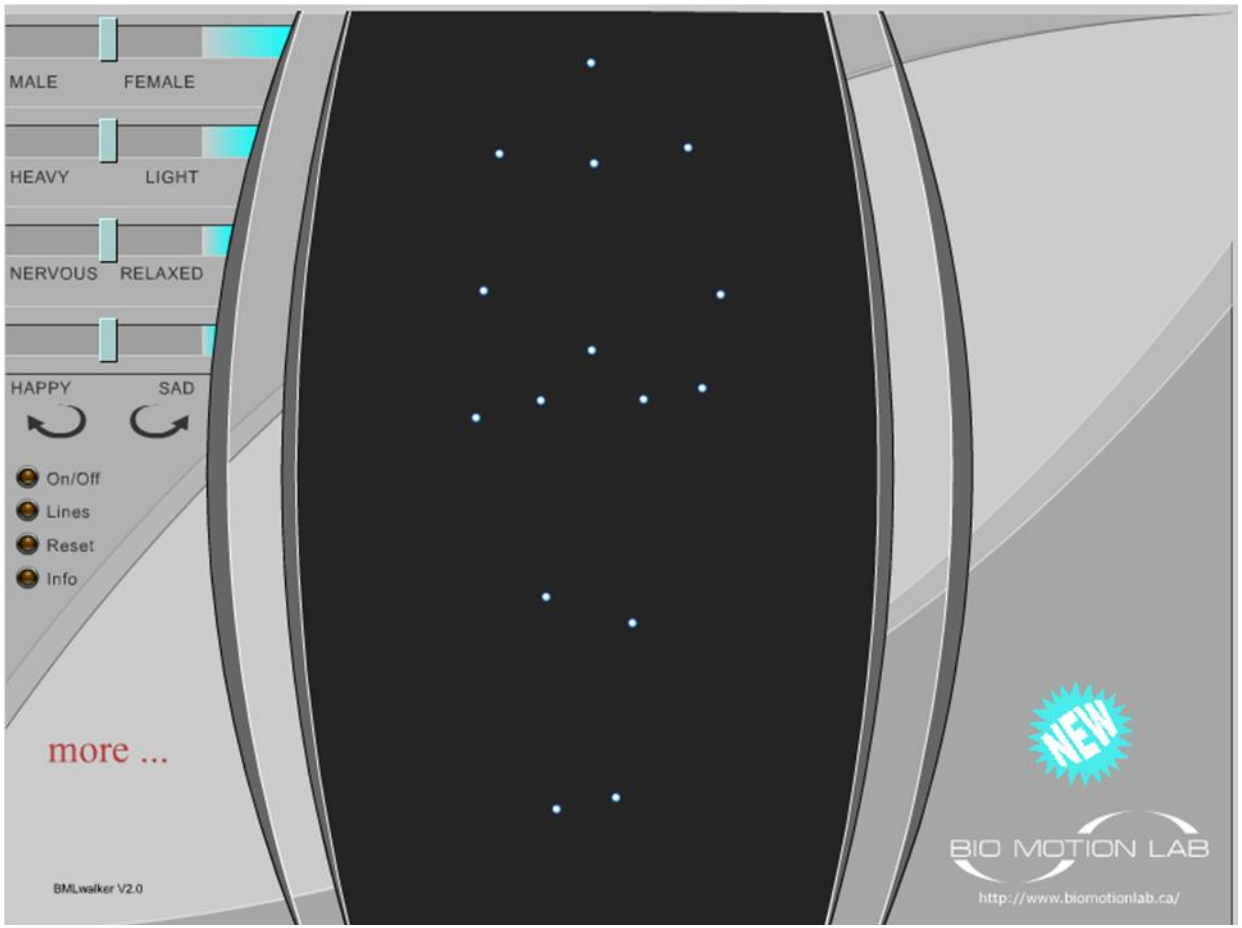
Движение



Что показано в видео?

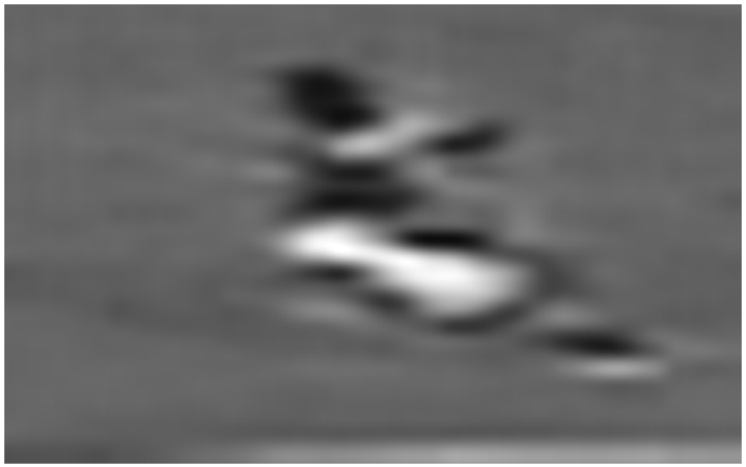
- Движение само по себе является мощной визуальной подсказкой
- Суть многих действий именно в динамике
- Иногда достаточно отследить движение отдельных точек, чтобы распознать событие

Распознавание по движению



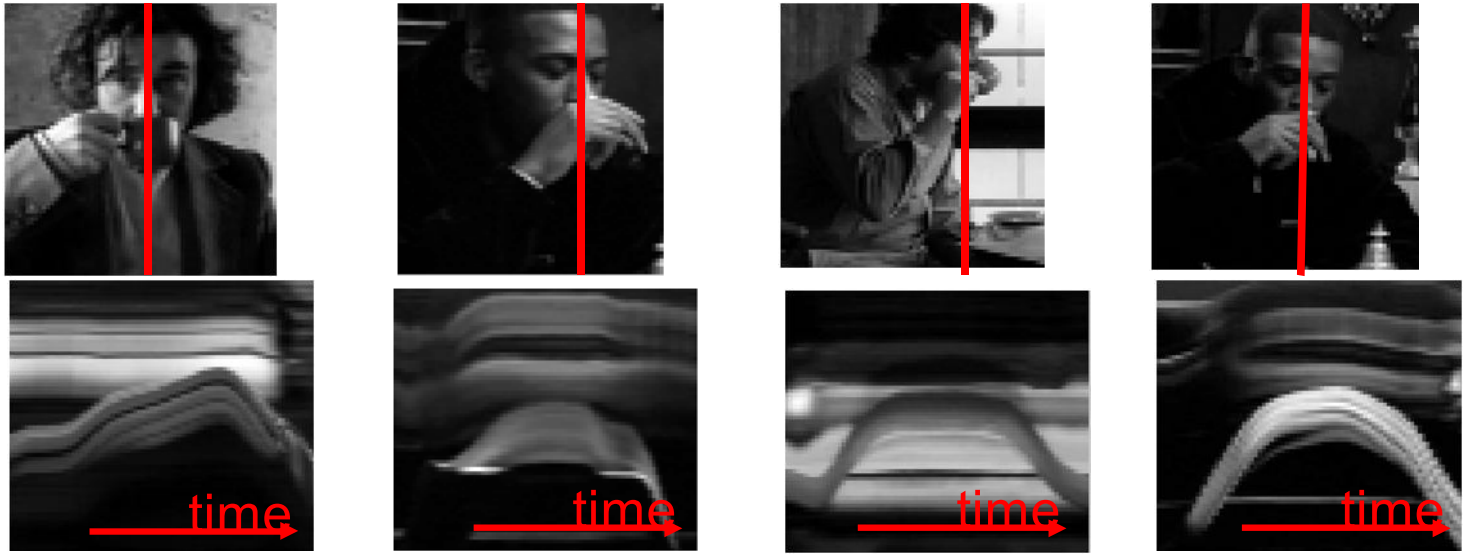
http://www.biomotionlab.ca/Demos/BMLwalker.html

Движение



- Одновременное описание изображением и движением позволяют распознавать события даже в низком разрешении

Скользящее окно



- Применим к видео схему «скользящее окно»
- Выделим фрагмент - пространственно-временной параллелепипед
- Возникают задачи, аналогичные задачам классификации изображений и поиска объектов, но для 3х мерного пространственно-временного объема

Microsoft Research

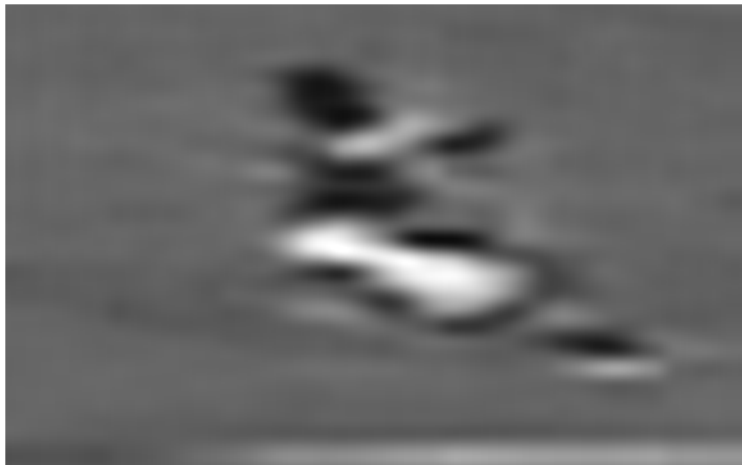
Задачи распознавания в видео

- Задачи, аналогичные задачам классификации изображений и поиска объектов, но для 3х мерного пространственно-временного объема
  - Классификация видеофрагмента
  - Поиск действий в видео (3D bbox)

- Можем применить наработанную методологию
  - Вычисление признаков (но уже по объему)
  - Поиск особых точек
  - Построение словарей
  - Классификацию
  - И т.д.

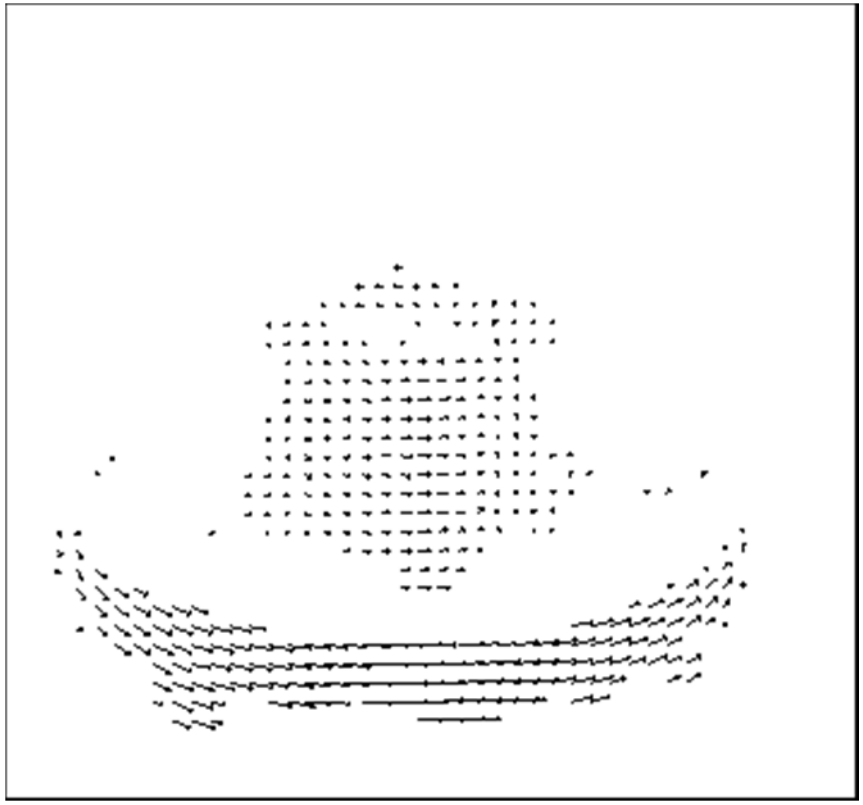
Microsoft Research

Описание движения



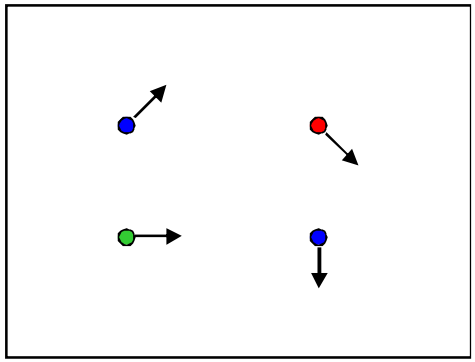
- Точки наблюдаемой сцены движутся относительно камеры / изображения
- Нужно это движение как-то формализовать, описывать и измерять

Оптический поток

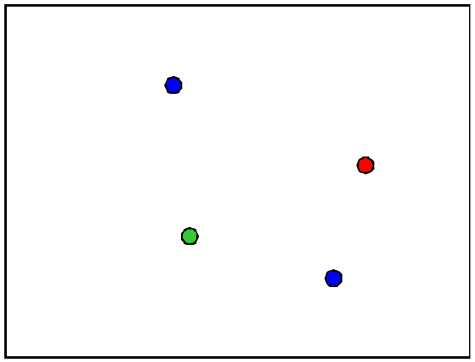


- Векторное поле движения пикселей между кадрами
- Задача - аналог задачи сопоставления изображений (dense matching)
- Один из базовых инструментов анализа изображений

Постановка задачи



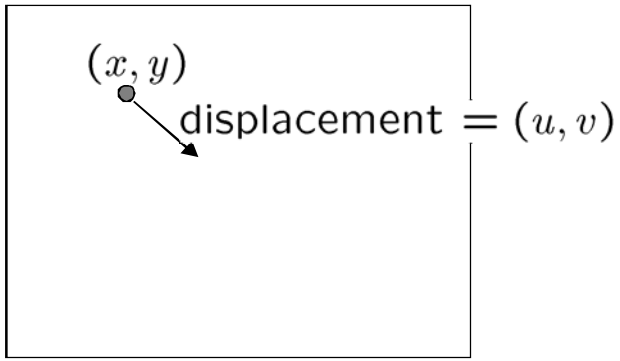
$H(x, y)$



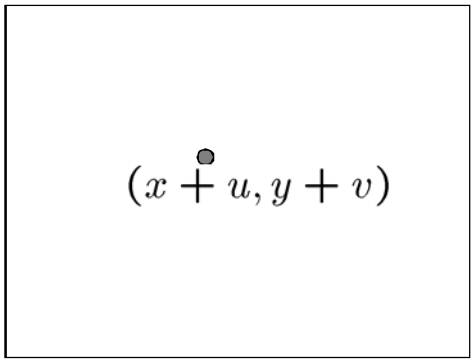
$I(x, y)$

- Как оценить движение пикселей от H в изображение I?
  - Проблема соответствия пикселей!
    - Пусть дан пиксель H, наити близкие пиксели того же цвета в I
- Ключевые предположения
  - Константный цвет: точка в H выглядит также, как и в I
    - Для изображения в градациях серого, это постоянная яркость
  - Малое движение: точки не уезжают далеко между кадрами

Ограничения на оптический поток



$H(x, y)$



$I(x, y)$

- Используем ограничения для формализации задачи
  - Постоянная яркость
  - Малое смещение: (u и v меньше 1-го пикселя)
    - Разложим функцию картинки в ряд тейлора I:

$$I(x+u, y+v) = I(x, y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \text{higher order terms}$$

$$\approx I(x, y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v$$

Microsoft Research

Уравнение оптического потока

- Объединим два ограничения

$$0 = I(x+u, y+v) - H(x,y)$$
$$\approx I(x,y) + I_x u + I_y v - H(x,y)$$
$$\approx (I(x,y) - H(x,y)) + I_x u + I_y v$$
$$\approx I_t + I_x u + I_y v$$
$$\approx I_t + \nabla I \cdot [u \ v]$$

$$I_x = \frac{\partial I}{\partial x}$$

В пределе u и v стремятся к нулю, и получаем равенство:

$$0 = I_t + \nabla I \cdot [\frac{\partial x}{\partial t} \ \frac{\partial y}{\partial t}]$$

Microsoft Research

Уравнение оптического потока

- Элементарное уравнение оптического потока:

$$0 = I_t + \nabla I \cdot [u \;\; v]$$

- Вопрос: сколько неизвестных и уравнений для каждого пикселя?
- 1 уравнение, 2 неизвестных (u, v)

Microsoft Research

Решение апертурной проблемы

- Как можно получить больше уравнений?
- Идея: наложить дополнительные ограничения
  - Пусть оптический поток меняется плавно
  - Вариант: пусть для всех пикселей *p* из окрестности (x,y) смещение (u,v) постоянно!
    - Для окна 5x5 получаем 25 уравнений для каждого пикселя!

$$0 = I_t(\mathbf{p_i}) + \nabla I(\mathbf{p_i}) \cdot [u\ v]$$

$$\underbrace{\begin{bmatrix} I_x(\mathbf{p_1}) & I_y(\mathbf{p_1}) \\ I_x(\mathbf{p_2}) & I_y(\mathbf{p_2}) \\ \vdots & \vdots \\ I_x(\mathbf{p_{25}}) & I_y(\mathbf{p_{25}}) \end{bmatrix}}_{A\ 25\times2} \underbrace{\begin{bmatrix} u \\ v \end{bmatrix}}_{d\ 2\times1} = - \underbrace{\begin{bmatrix} I_t(\mathbf{p_1}) \\ I_t(\mathbf{p_2}) \\ \vdots \\ I_t(\mathbf{p_{25}}) \end{bmatrix}}_{b\ 25\times1}$$

Microsoft Research

Цвет вместо яркости

- При использовании окна 5x5 получается 25*3 уравнений на пиксель!

$$0 = I_t(\mathbf{p_i})[0,1,2] + \nabla I(\mathbf{p_i})[0,1,2] \cdot [u\ v]$$

$$\underbrace{\begin{bmatrix} I_x(\mathbf{p_1})[0] & I_y(\mathbf{p_1})[0] \\ I_x(\mathbf{p_1})[1] & I_y(\mathbf{p_1})[1] \\ I_x(\mathbf{p_1})[2] & I_y(\mathbf{p_1})[2] \\ \vdots & \vdots \\ I_x(\mathbf{p_{25}})[0] & I_y(\mathbf{p_{25}})[0] \\ I_x(\mathbf{p_{25}})[1] & I_y(\mathbf{p_{25}})[1] \\ I_x(\mathbf{p_{25}})[2] & I_y(\mathbf{p_{25}})[2] \end{bmatrix}}_{A\ 75\text{x}2} \underbrace{\begin{bmatrix} u \\ v \end{bmatrix}}_{d\ 2\text{x}1} = -\underbrace{\begin{bmatrix} I_t(\mathbf{p_1})[0] \\ I_t(\mathbf{p_1})[1] \\ I_t(\mathbf{p_1})[2] \\ \vdots \\ I_t(\mathbf{p_{25}})[0] \\ I_t(\mathbf{p_{25}})[1] \\ I_t(\mathbf{p_{25}})[2] \end{bmatrix}}_{b\ 75\text{x}1}$$

Microsoft Research

Алгоритм Лукаса-Канаде

- Проблема: больше уравнений, чем неизвестных!

$$\underset{25\text{x}2}{A}\ \underset{2\text{x}1}{d} = \underset{25\text{x}1}{b} \longrightarrow \text{minimize } \|Ad - b\|^2$$

- Получаем задачу наименьших квадратов
- Можем решить её через нормальные уравнения

$$\underset{2\text{x}2}{(A^T A)}\ \underset{2\text{x}1}{d} = \underset{2\text{x}1}{A^T b} \Rightarrow d = (A^T A)^{-1} A^T b$$

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

$A^T A$ $\qquad$ $A^T b$

- Суммируем по всем пикселям в окне K x K
- Это метод был предложен Лукасом и Канаде в 1981 году

Microsoft Research

Алгоритм Лукаса-Канаде

- Оптимальные (u, v) удовлетворяют уравнению

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

$$A^T A \qquad\qquad A^T b$$

- Которое может быть решено через нормальные уравнения

$$d = (A^T A)^{-1} A^T b$$

B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *DARPA Image Understanding Workshop, April 1981.*

Microsoft Research

Условия на разрешимость

- Решение задачи оптического потока d = (u,v) может быть найдено в виде

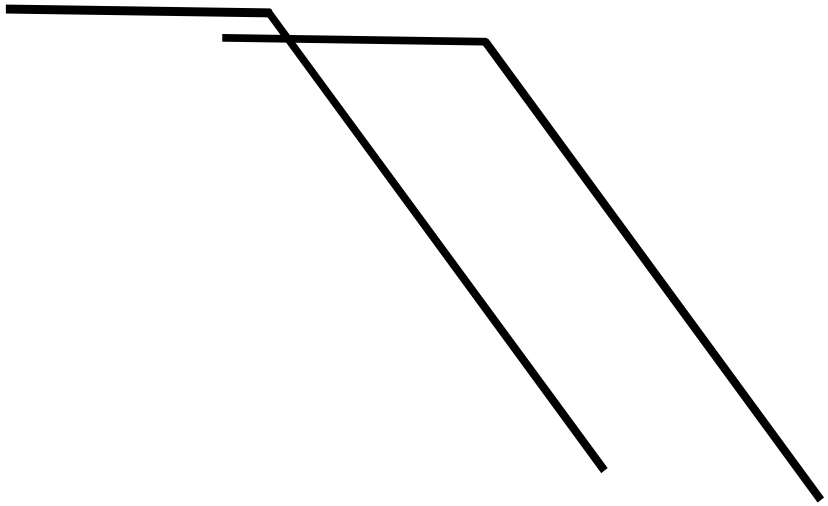
$$A^T A = \begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} = \sum \begin{bmatrix} I_x \\ I_y \end{bmatrix} [I_x \; I_y] = \sum \nabla I (\nabla I)^T$$

$$d = (A^T A)^{-1} A^T b \qquad A^T b = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

- Когда задача разрешима?
  - A^TA должна быть обратимой
  - A^TA не должна быть слишком близка к нулю
    - С.значения λ_1 и λ_2 матрицы A^TA не должны быть малы
  - A^TA должна быть хорошо определима
    - λ_1 / λ_2 не должно быть слишком велико
    - (λ_1 = наибольшее с.значение)
- A^TA разрешима, когда нет апертурной проблемы

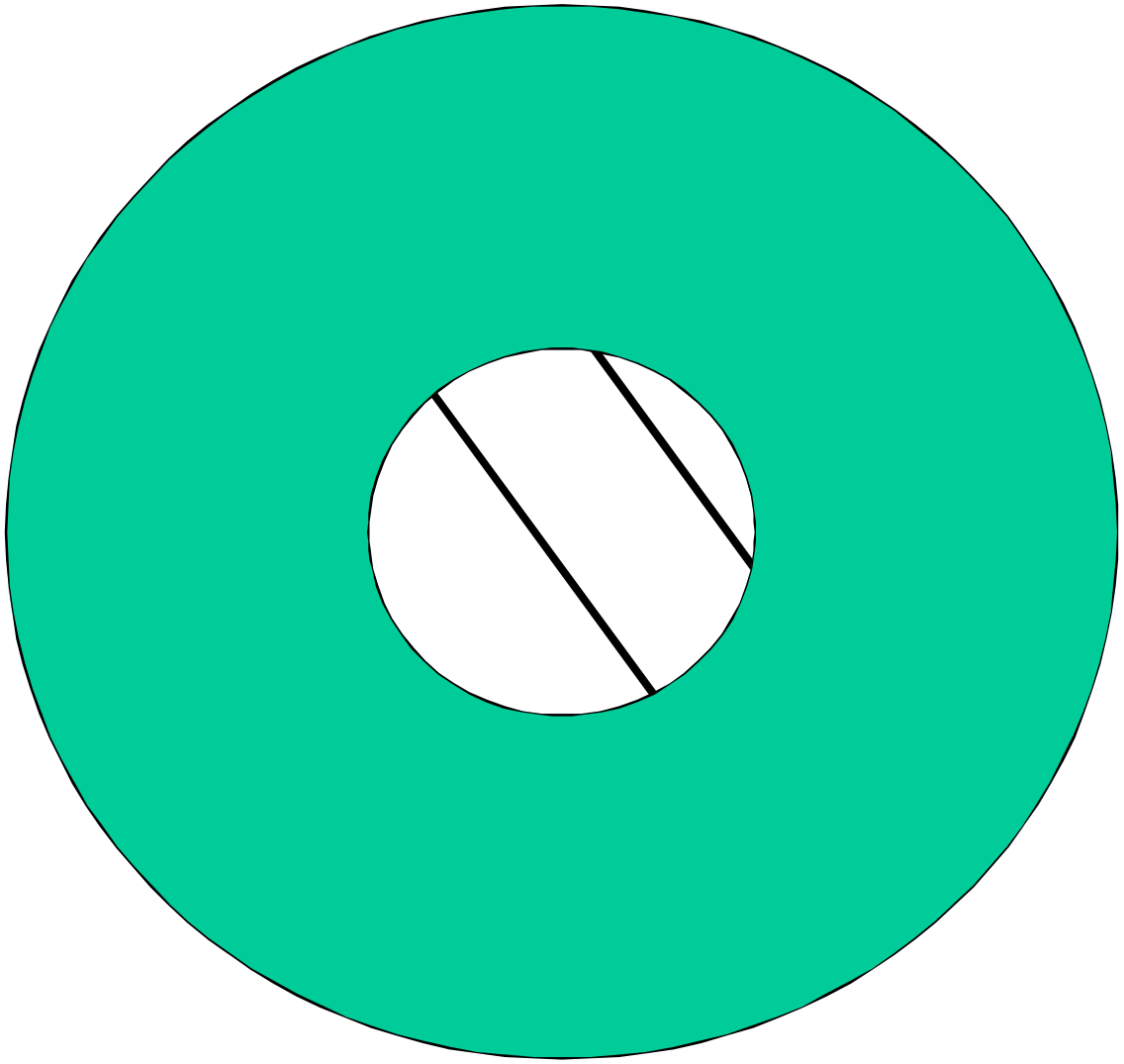
Проблема апертуры

Microsoft Research

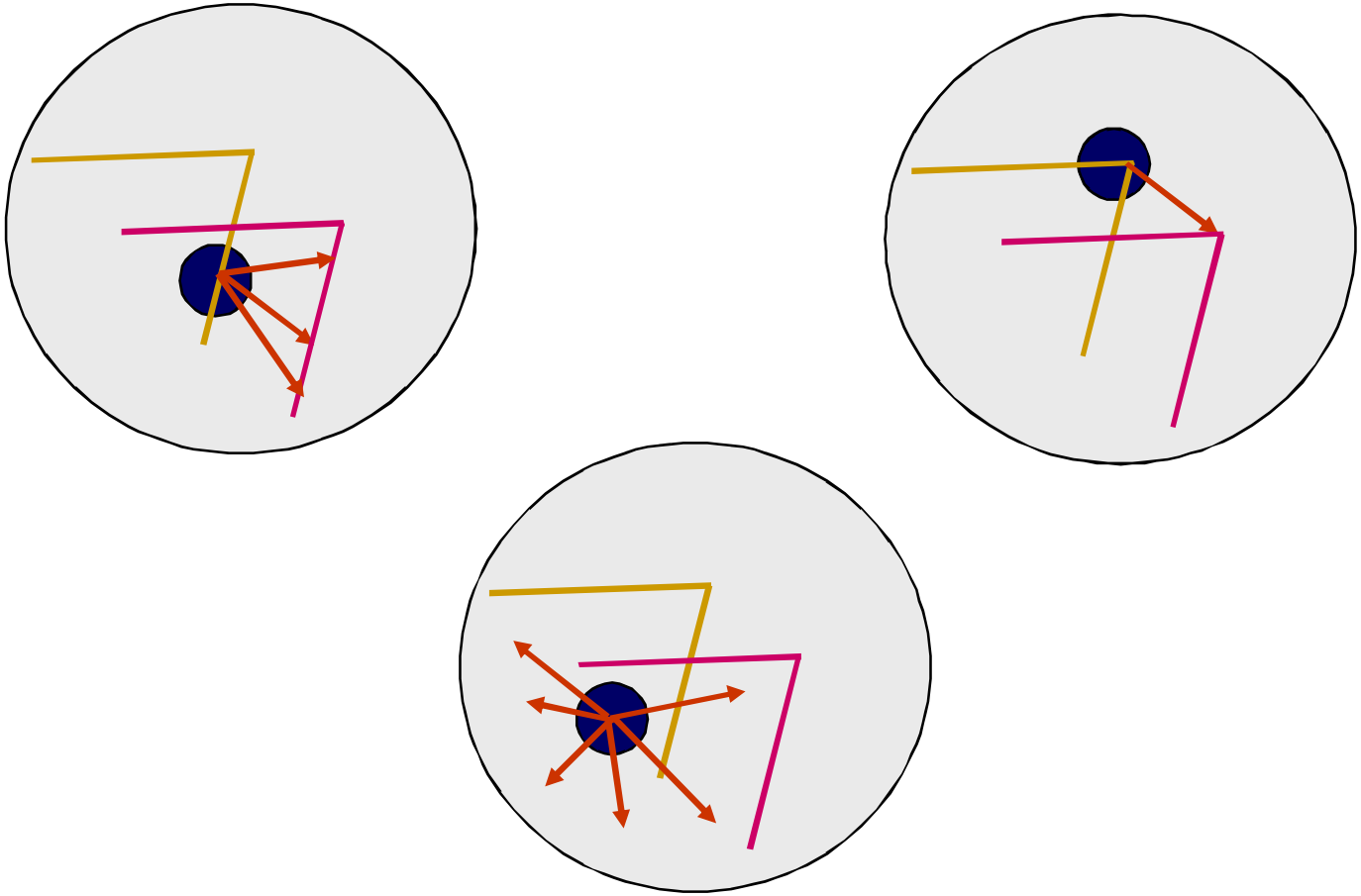


Проблема апертуры

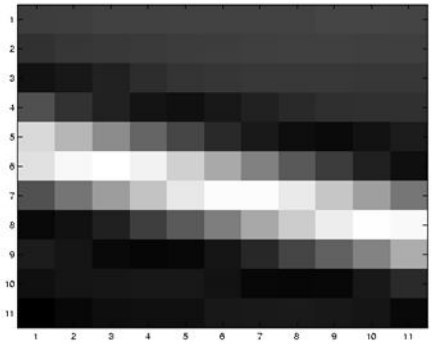
Microsoft Research

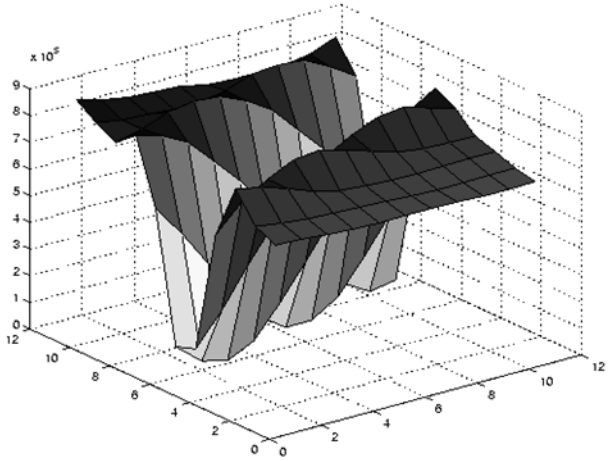


Анализ участка изображения



Края



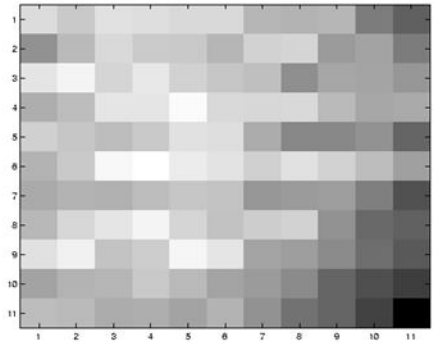


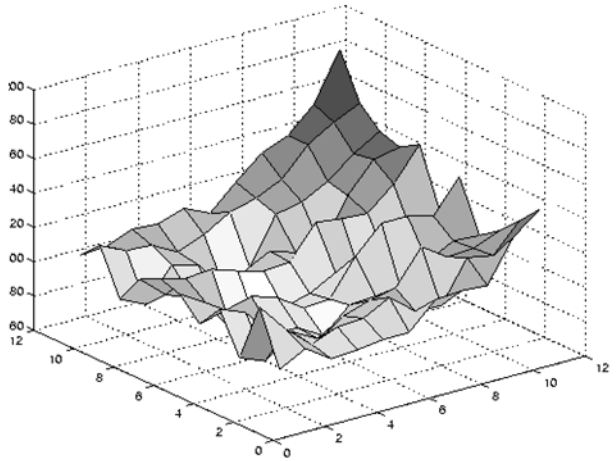
$$\sum \nabla I (\nabla I)^T$$

– большие градиенты

– большое λ_1, маленькое I λ_2

Слабоконтрастная текстура



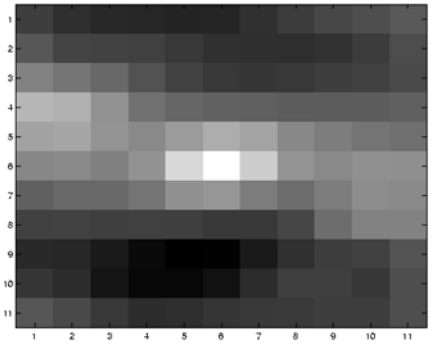


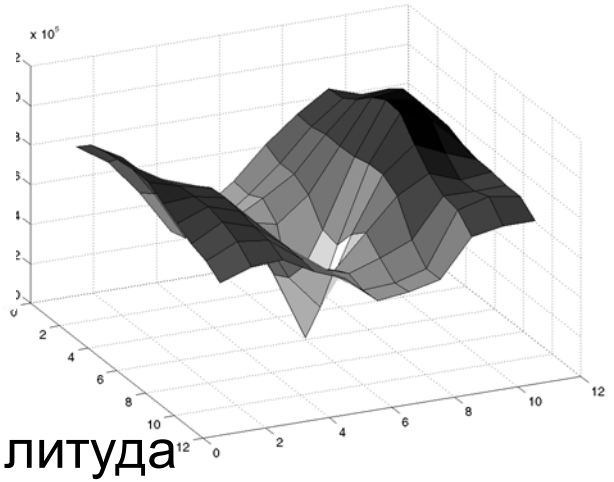
$$\sum \nabla I (\nabla I)^T$$

– величина градиента мала

– малое λ_1, малое λ_2

Текстурированная область





$$\sum \nabla I (\nabla I)^T$$

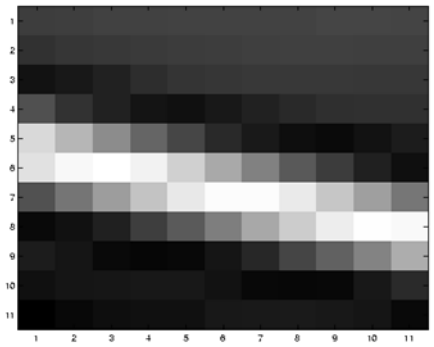
– градиенты разные, большая амплитуда

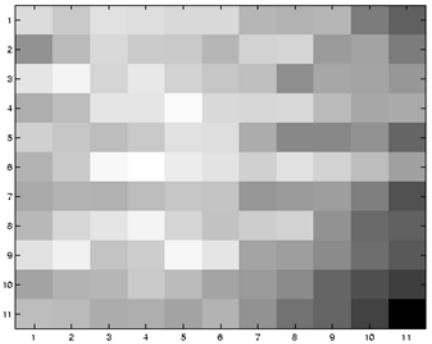
– большое λ_1, большое λ_2

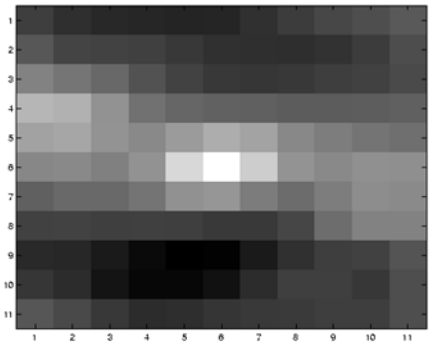
Microsoft Research

Замечание

- Поиск оптического потока производится между двумя изображениями, но:
  - Можем оценить качество оценки только по 1 изображению!
  - По 1 картинке можно сказать, в каких точках поток будет считаться хорошо, а в каких - нет
  - На этом основаны методы выбора особенностей для отслеживания
  - Фактически, поиск особых точек!







Jianbo Shi and Carlo Tomasi, "Good Features to Track," *CVPR 1994*

Microsoft Research

Погрешности метода Л-К

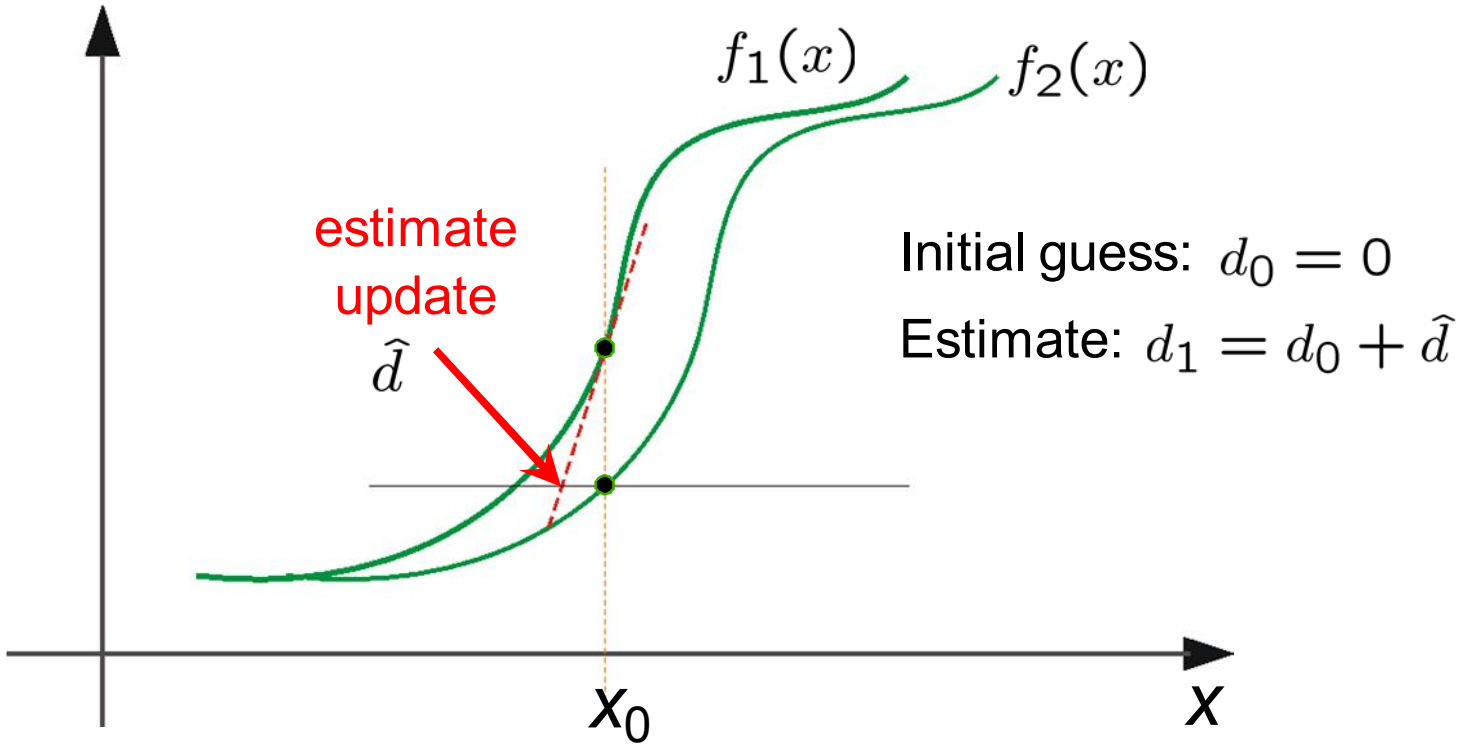
- Каковы потенциальные источники ошибок?
  - Предполагаем, что A^TA обратима
  - Предполагаем, что в изображении мало шума
- Когда эти предположения нарушаются:
  - Яркость точки не постоянная
  - Движение между кадрами большое
  - Движение соседей отличается от движения точки
    - Окно поиска слишком большое
    - Какой наилучший размер окна поиска?

Microsoft Research

Итеративное уточнение

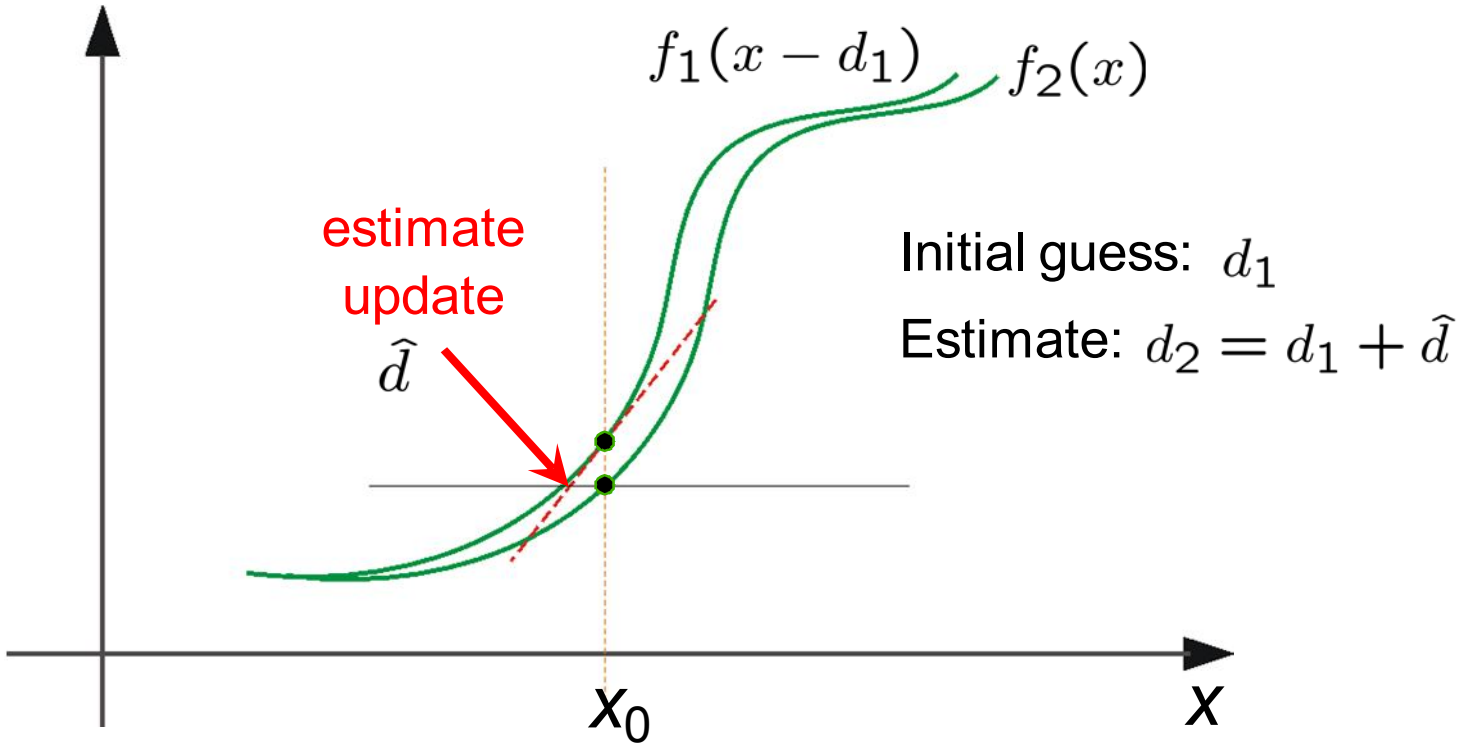
- Итеративный вариант алгоритма Лукаса-Канаде
  1. Оценить движение в каждом пикселе, решив уравнения Лукаса-Канаде
  2. Преобразовать изображение H используя вычисленное движение
  3. Повторить 1-2 до сходимости

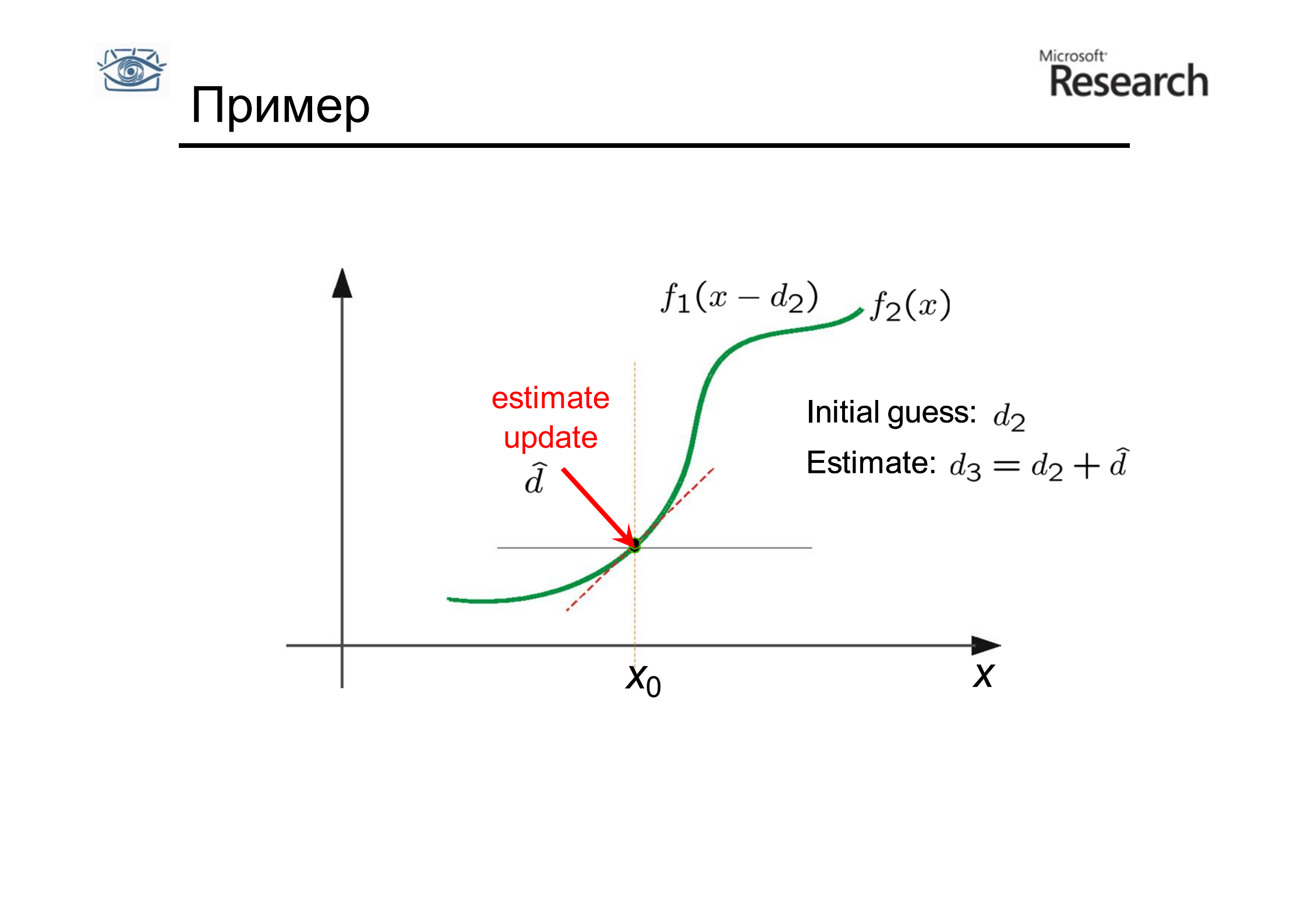
Пример



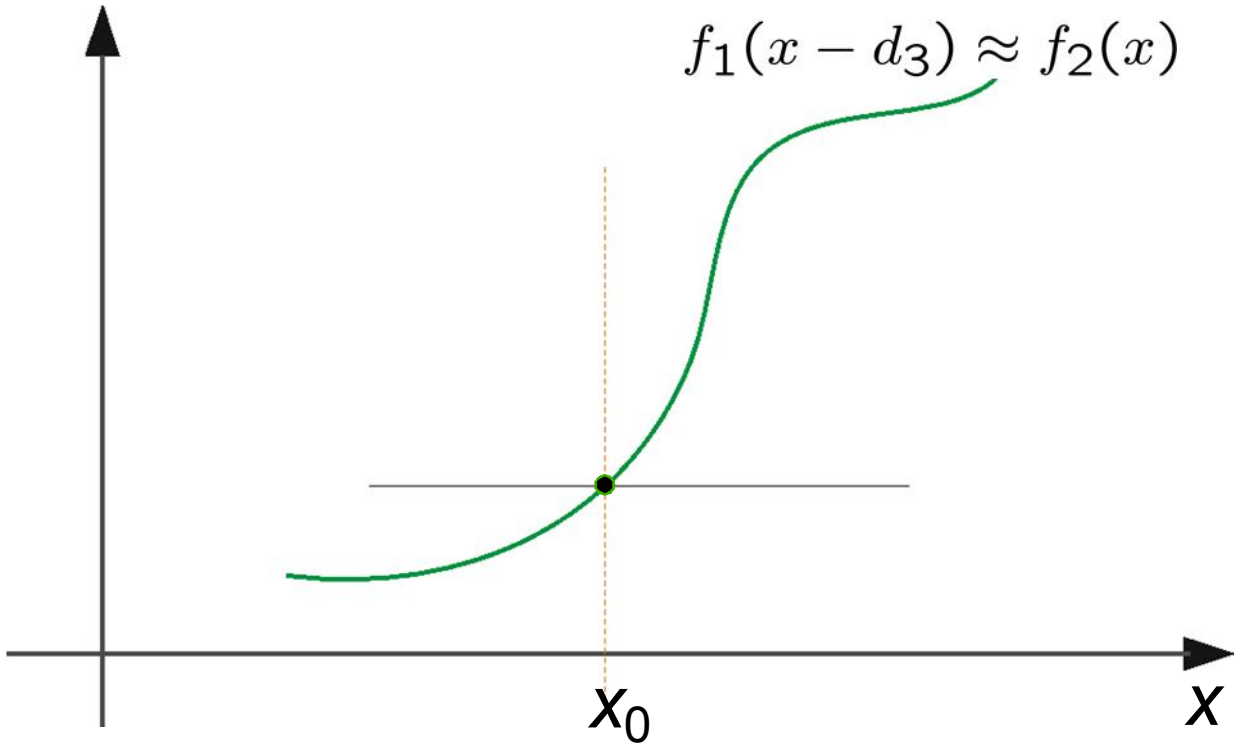
$0 = I_t + I_x d$ - Рассмотрим одномерный случай (d – смещение)

Пример


$f_1(x - d_1)$
$f_2(x)$
estimate
update
$\hat{d}$
Initial guess: d_1
Estimate: $d_2 = d_1 + \hat{d}$
x_0
x


Microsoft Research
Пример
$f_1(x - d_2)$
$f_2(x)$
estimate
update
$\hat{d}$
Initial guess: d_2
Estimate: $d_3 = d_2 + \hat{d}$
x_0
x

Пример



Microsoft Research

Замечания

- Сглаживаем изображения для более аккуратного вычисления градиентов
- Лучше градиенты вычислять по одному изображению, а преобразовывать другое
- Преобразование изображения приводит к ошибкам из-за дискретизации, можем потерять в точности

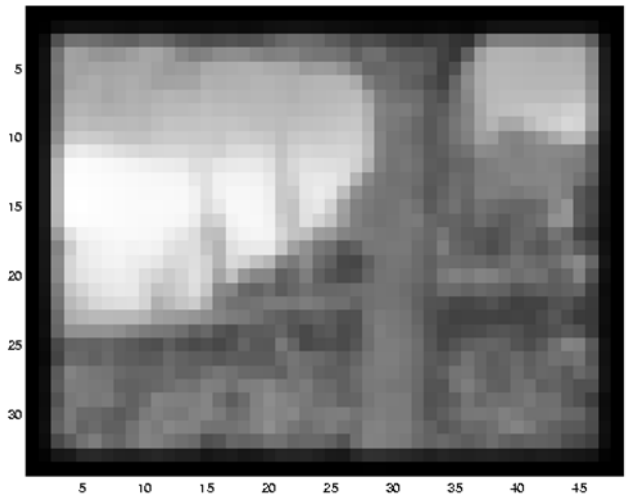
Проблема большого смещения

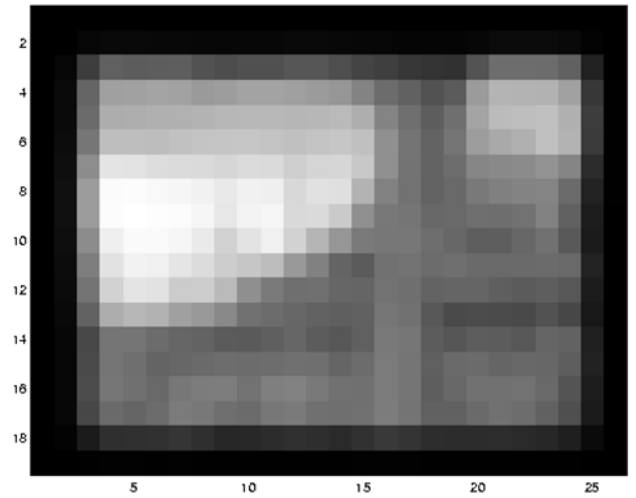
Microsoft Research

- Насколько мало движение в изображении?
  - Существенно больше 1-го пикселя
  - Как можно решить эту проблему?

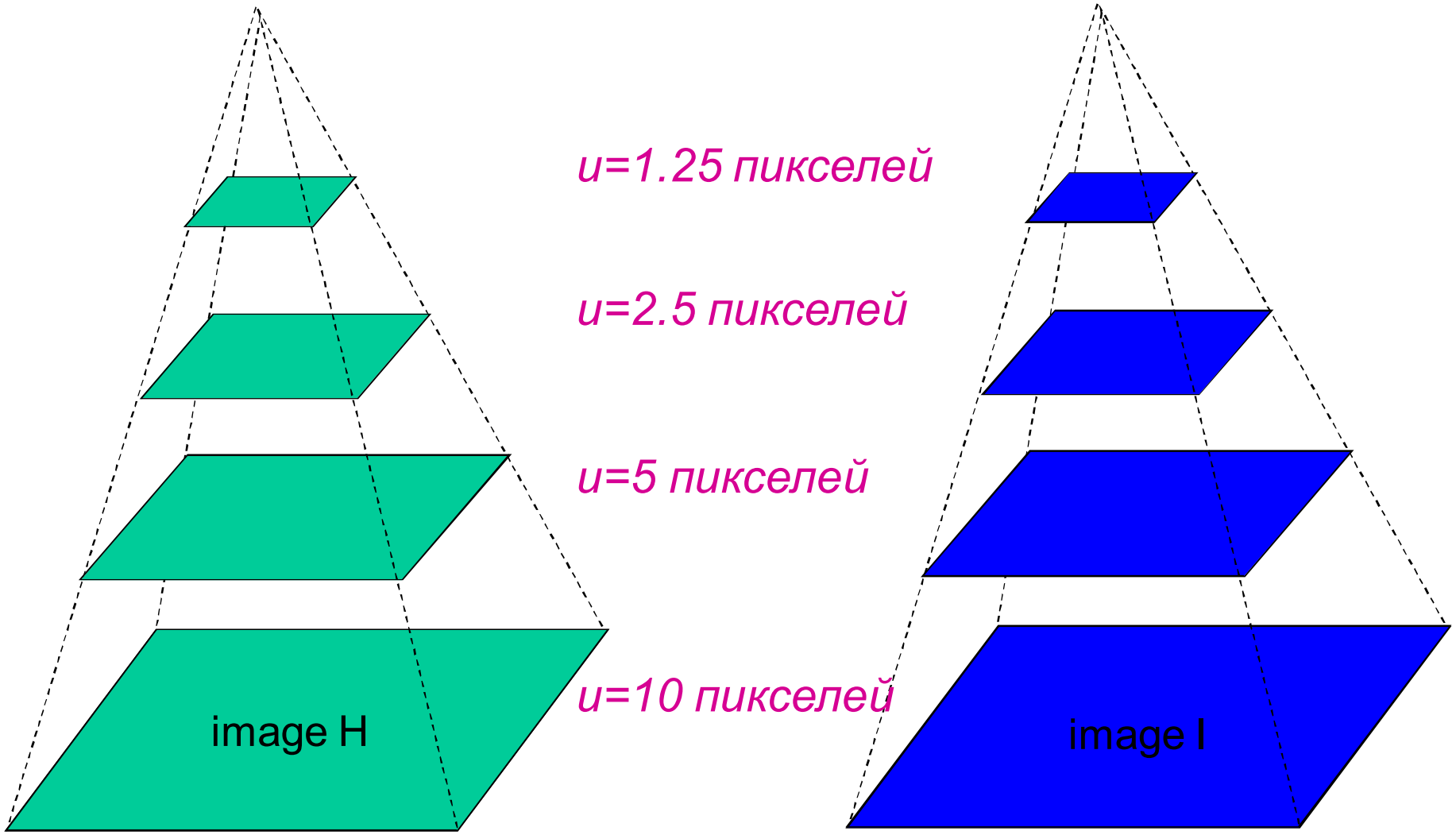
Пирамида разрешений

Microsoft Research





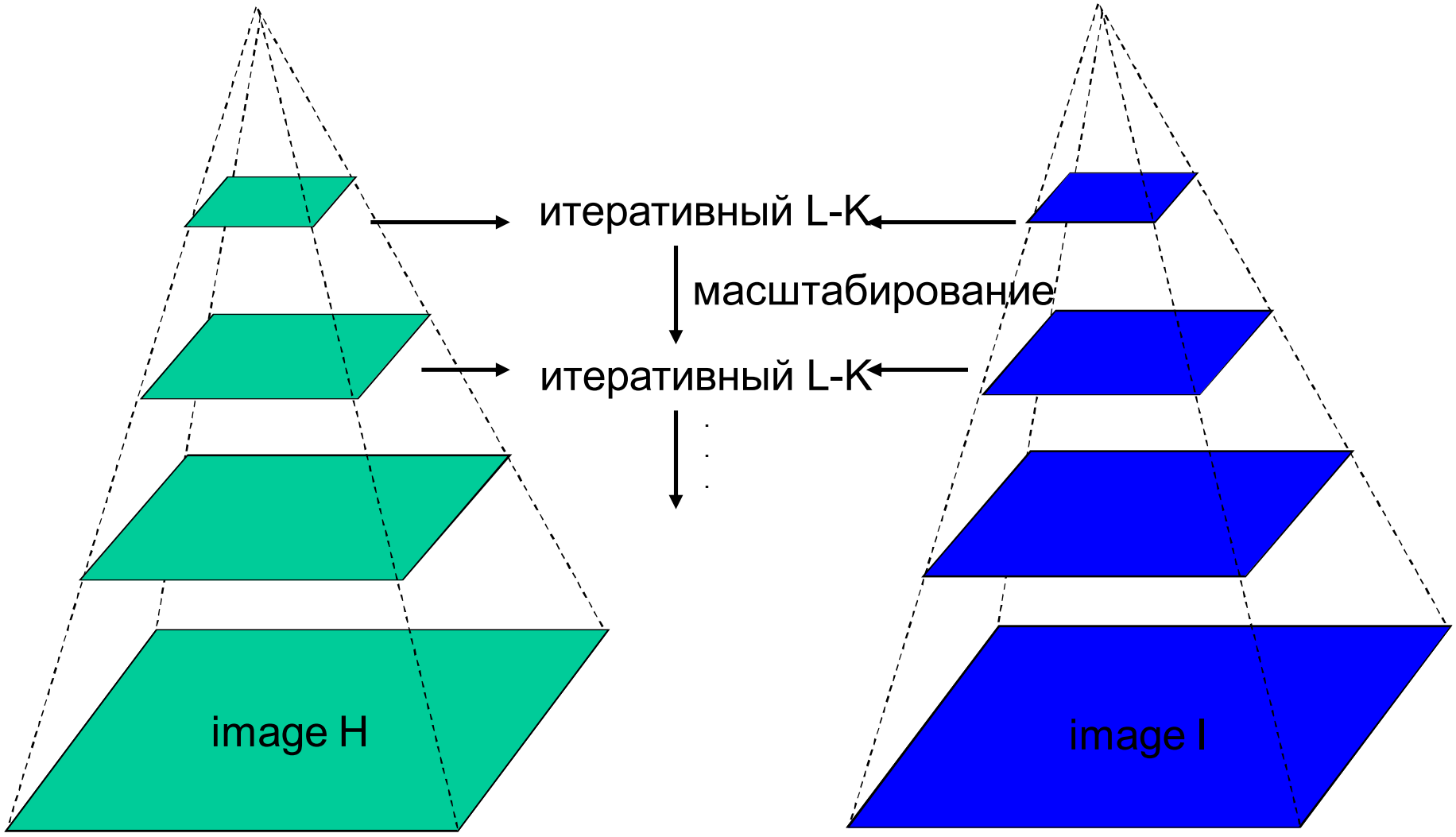
Иерархический метод



Гауссова пирамида для H

Гауссова пирамида для I

Иерархический метод



Гауссова пирамида для H

Гауссова пирамида для I

Microsoft Research

Другие модели движения

- В рассматриваемых алгоритмах модель движение – параллельный перенос (u,v)
- Можно использовать другие модели
  - поворот, аффинную, перспективную
- Необходимо просто вычислить соответствующий Якобиан

$$\mathbf{A}^{\mathbf{T}}\mathbf{A} = \sum_i \mathbf{J}\nabla\mathbf{I}(\nabla\mathbf{I})^{\mathbf{T}}\mathbf{J}^{\mathbf{T}}$$

$$\mathbf{A}^{\mathbf{T}}\mathbf{b} = -\sum_i \mathbf{J}^{\mathbf{T}} I_t (\nabla\mathbf{I})^{\mathbf{T}}$$

Проблемы:

- Много параметров, столько же исходных данных
- Меньшая устойчивость и надежность

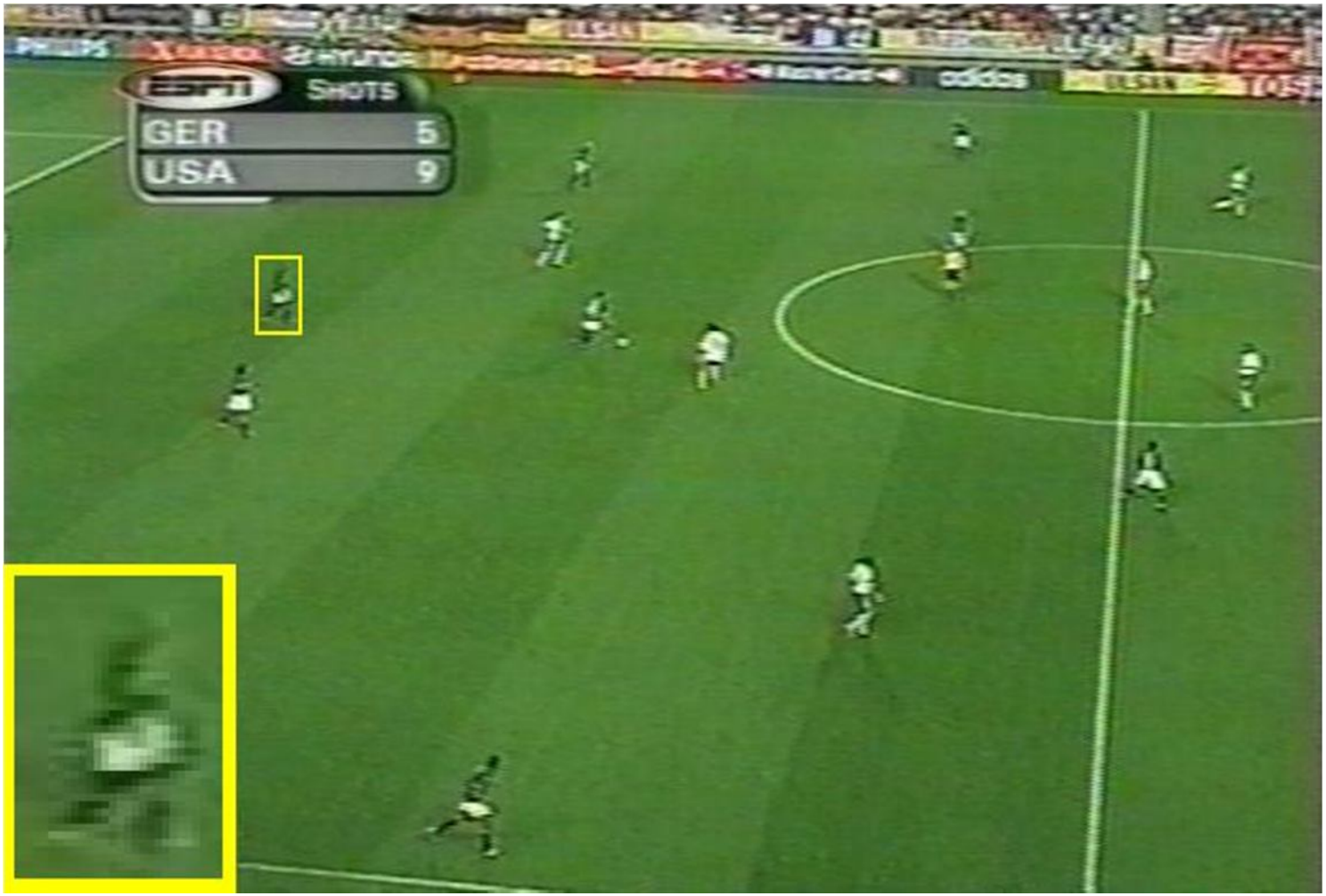
Jianbo Shi and Carlo Tomasi, "Good Features to Track," *CVPR 1994*

Microsoft Research

Реализации

- OpenCV
  - GoodFeatureToTrack
    - Выбор особых точек (фактически, Harris)
  - cvCalcOpticalFlowPyrLK
    - Иерархическое расширение метода Lucas-Kanade для оптического потока

Распознавание событий



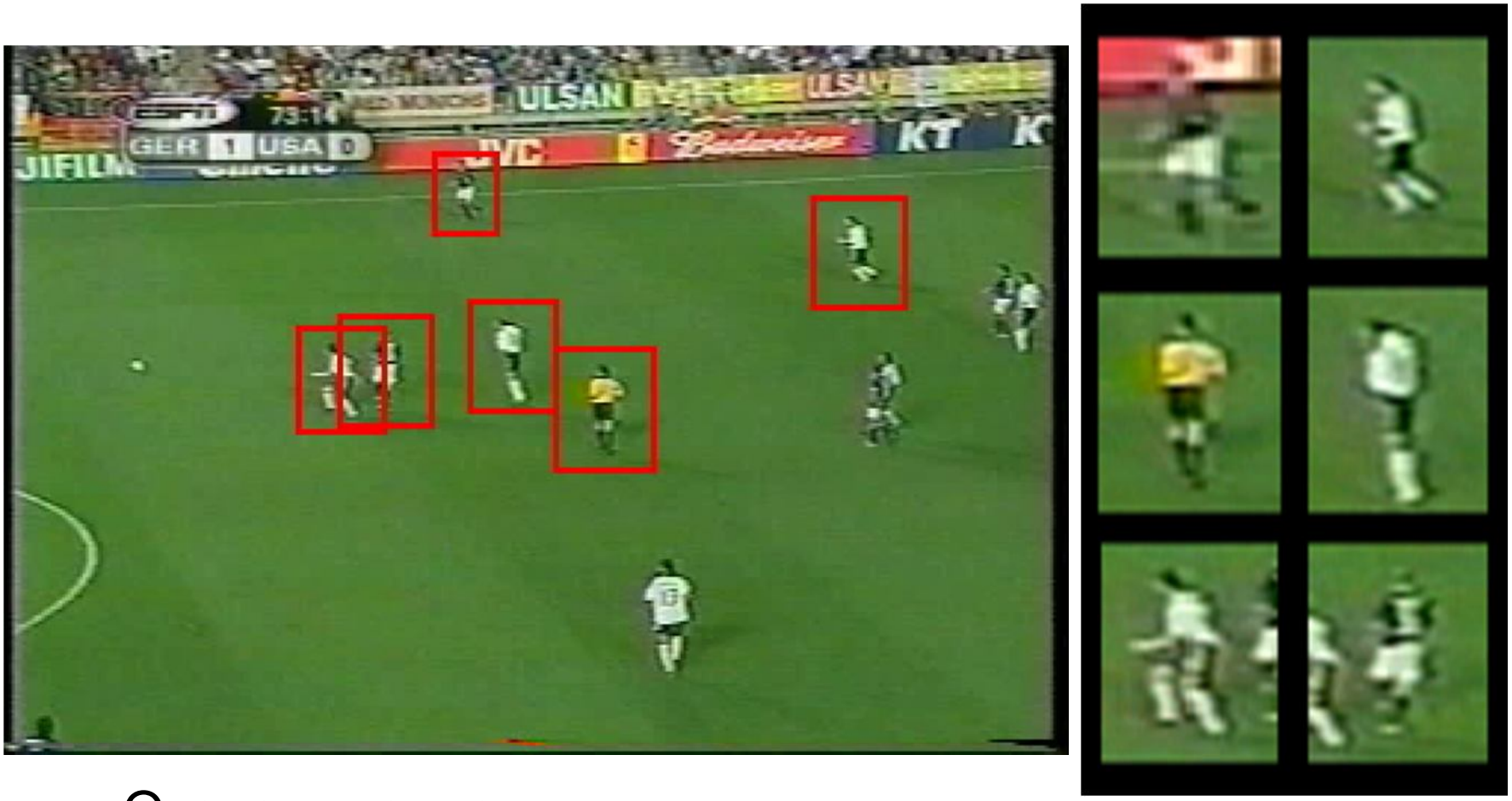
30и пиксельный человечек

Alexei A. Efros, Alexander C. Berg, Greg Mori and Jitendra Malik. Recognizing Action at a Distance. ICCV 2003

Задача

- Распознавать события (действия человека) на расстоянии
  - Низкое разрешение, шумные данные
  - Движущаяся камера, перекрытия
  - Широкая гамма событий

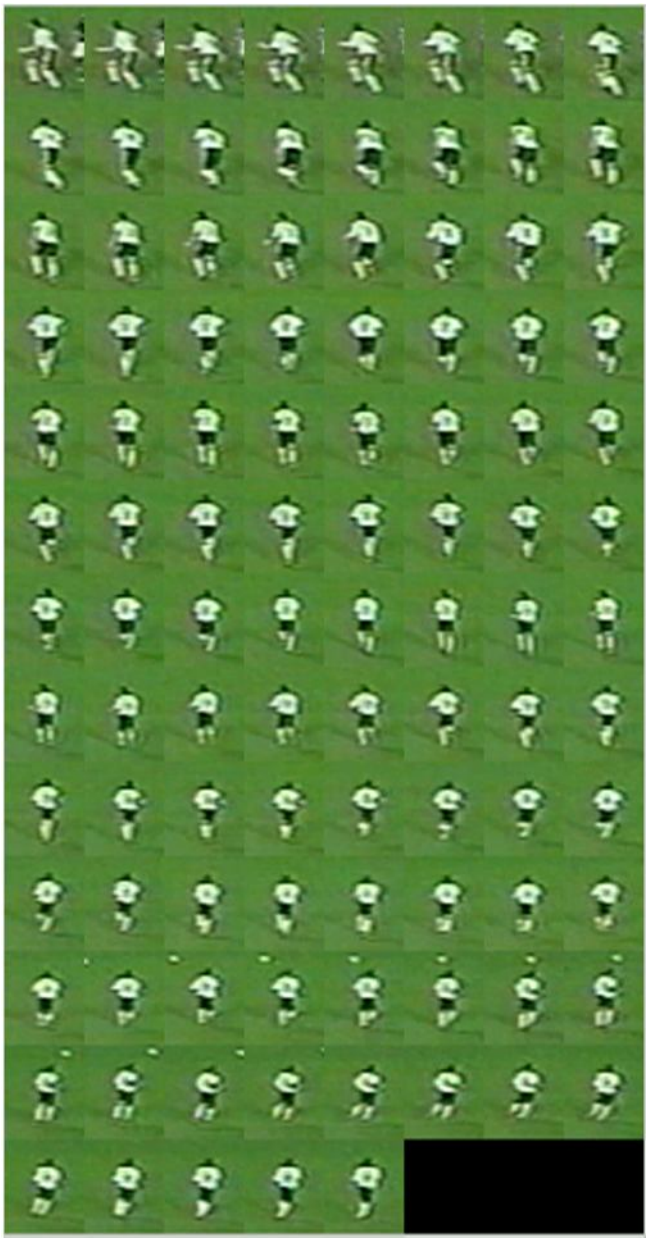
Сбор данных



- Отслеживание
  - Инициализируется пользователем
  - Простой коррелятор

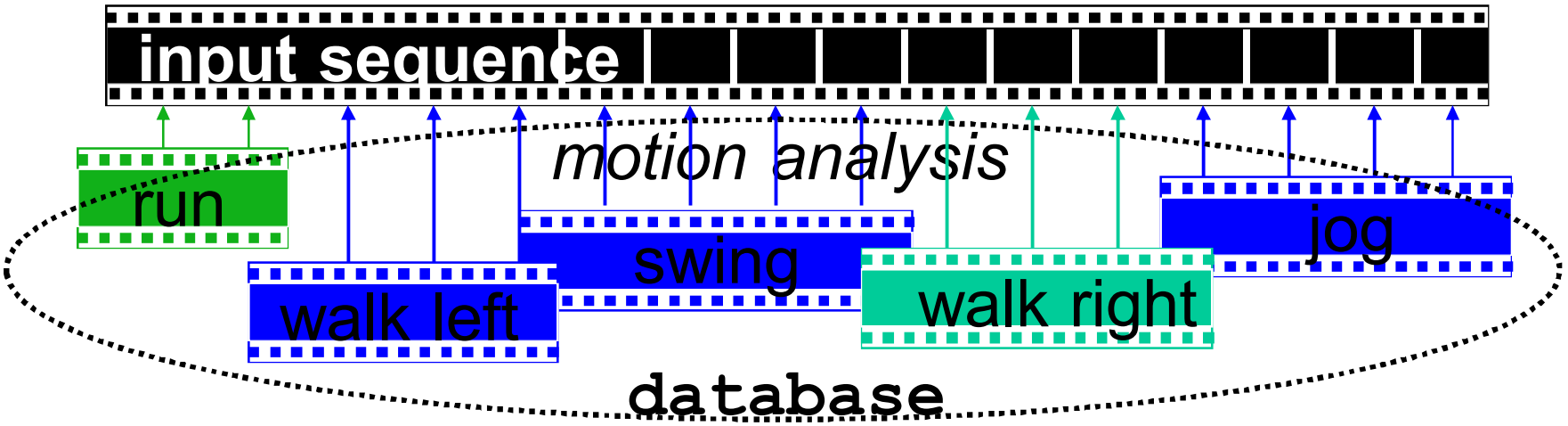
Представление

- Стабилизированный пространственно-временной объем
  - Нет информации о перемещении
  - Всё движение вызвано движением конечностей человека

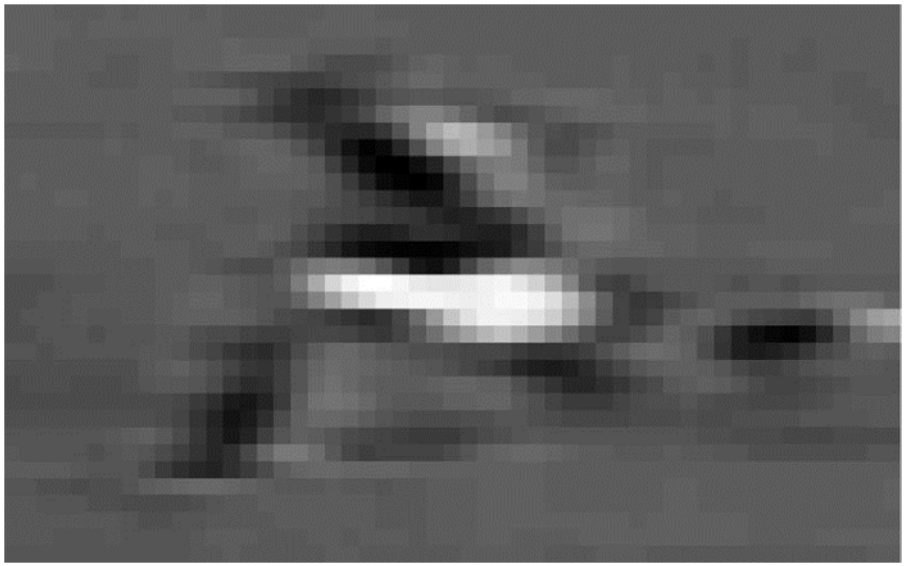


Схема

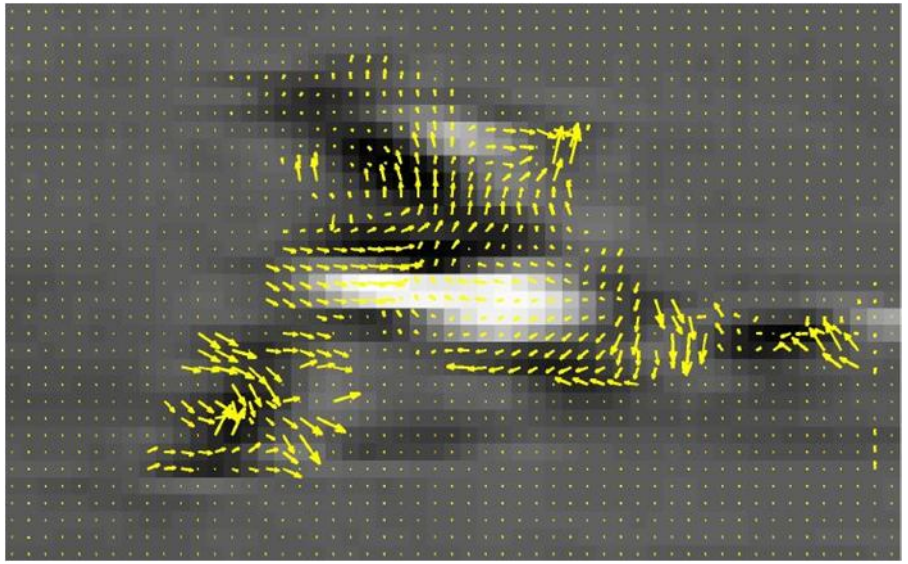
- Будем аннотировать видеопоток сопоставляя его с ранее записанными видеофрагментами
  - Для каждого кадра, сопоставим фрагмент в некотором временном интервале
  - Распознавание через сопоставление (Nearest Neighbor)



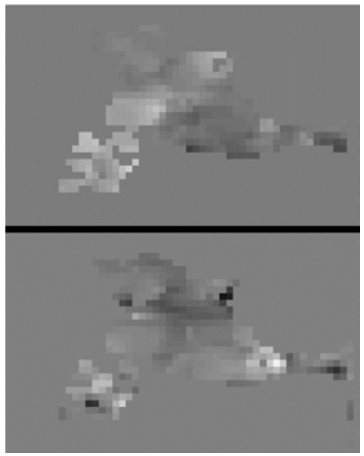
Дескриптор движения



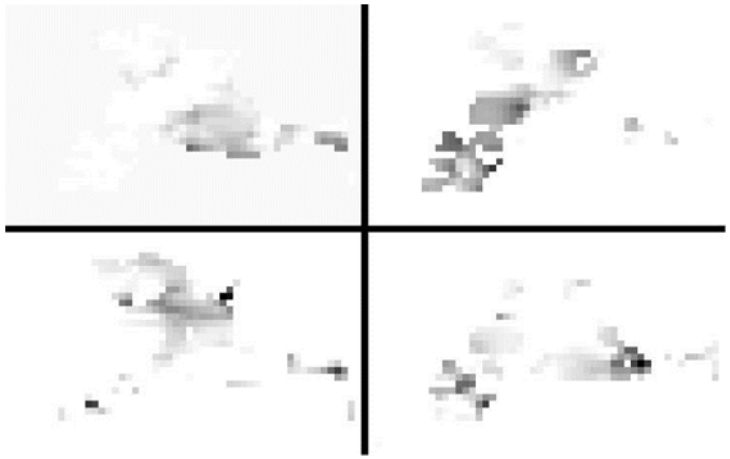
Изображение



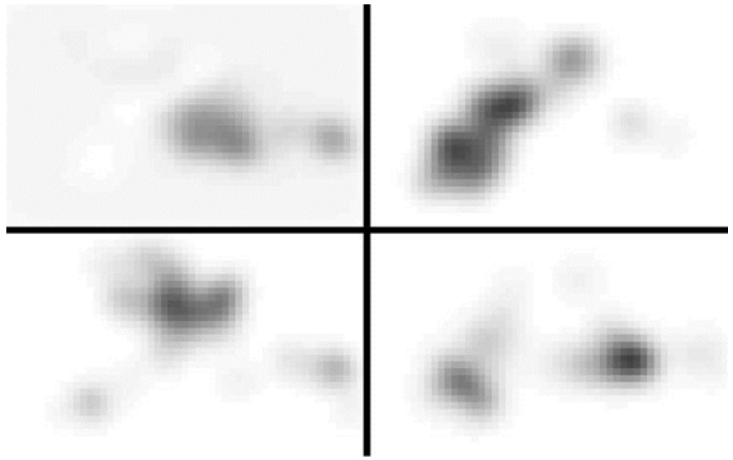
Оптический поток $F_{x,y}$



F_x, F_y

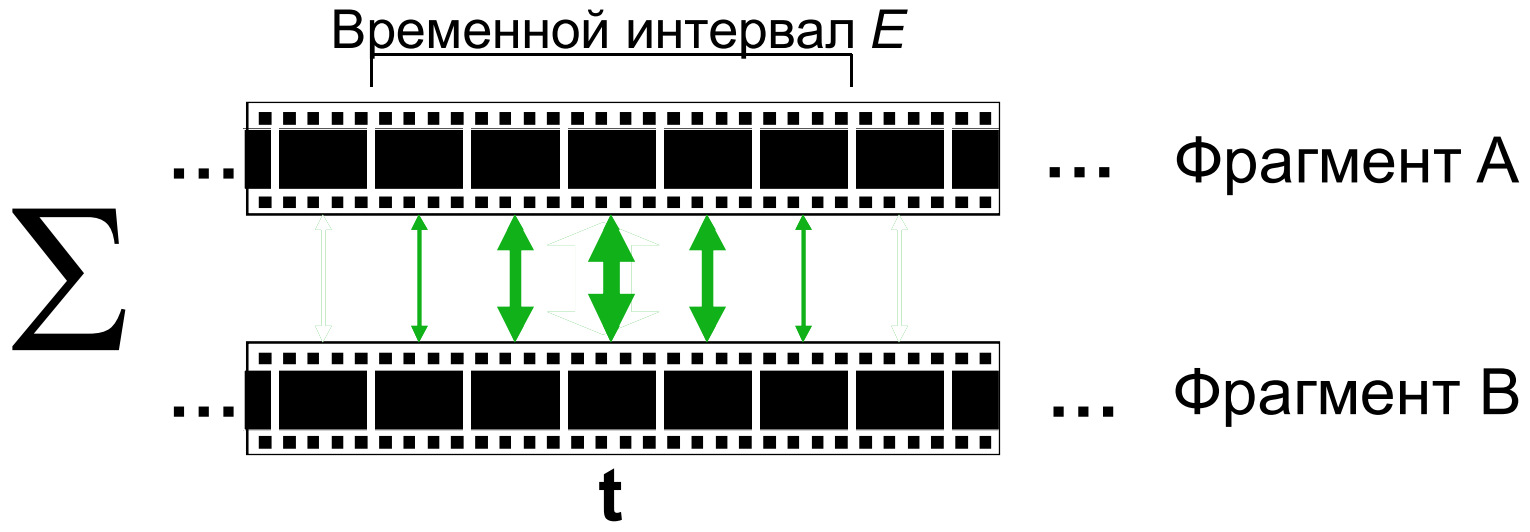


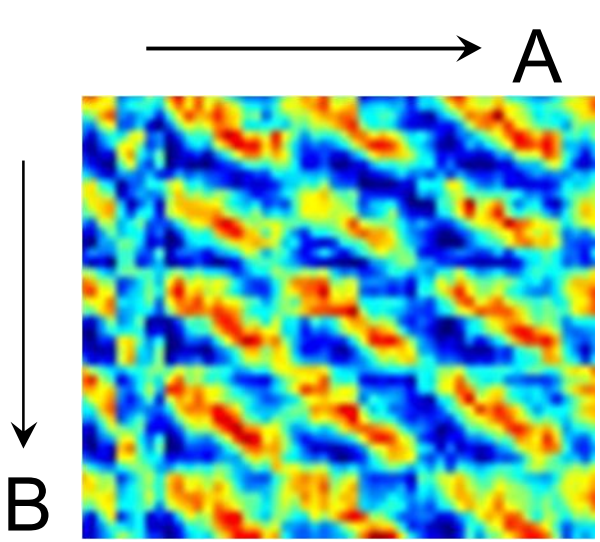
$F_x^-, F_x^+, F_y^-, F_y^+$



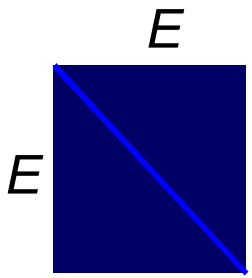
сглаженные $F_x^-, F_x^+, F_y^-, F_y^+$

Дескриптор движения

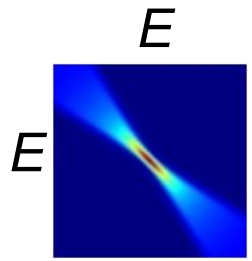




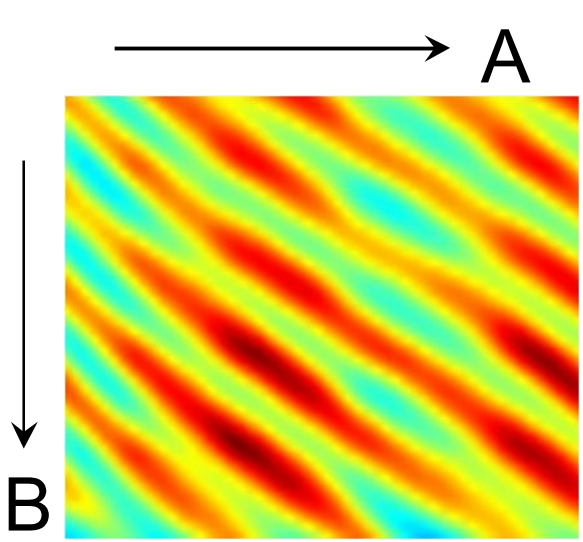
Покадровая
Матрица сходства



I -матрица



сглаженная I

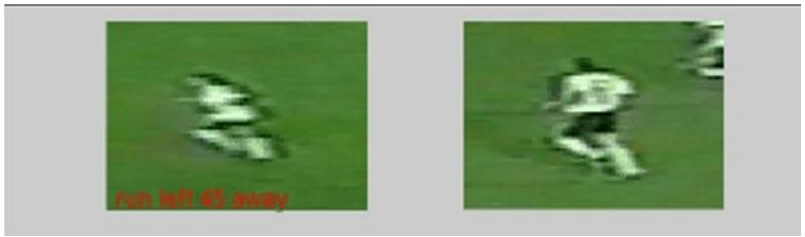


Матрица сходства
движений

Действия: сопоставление

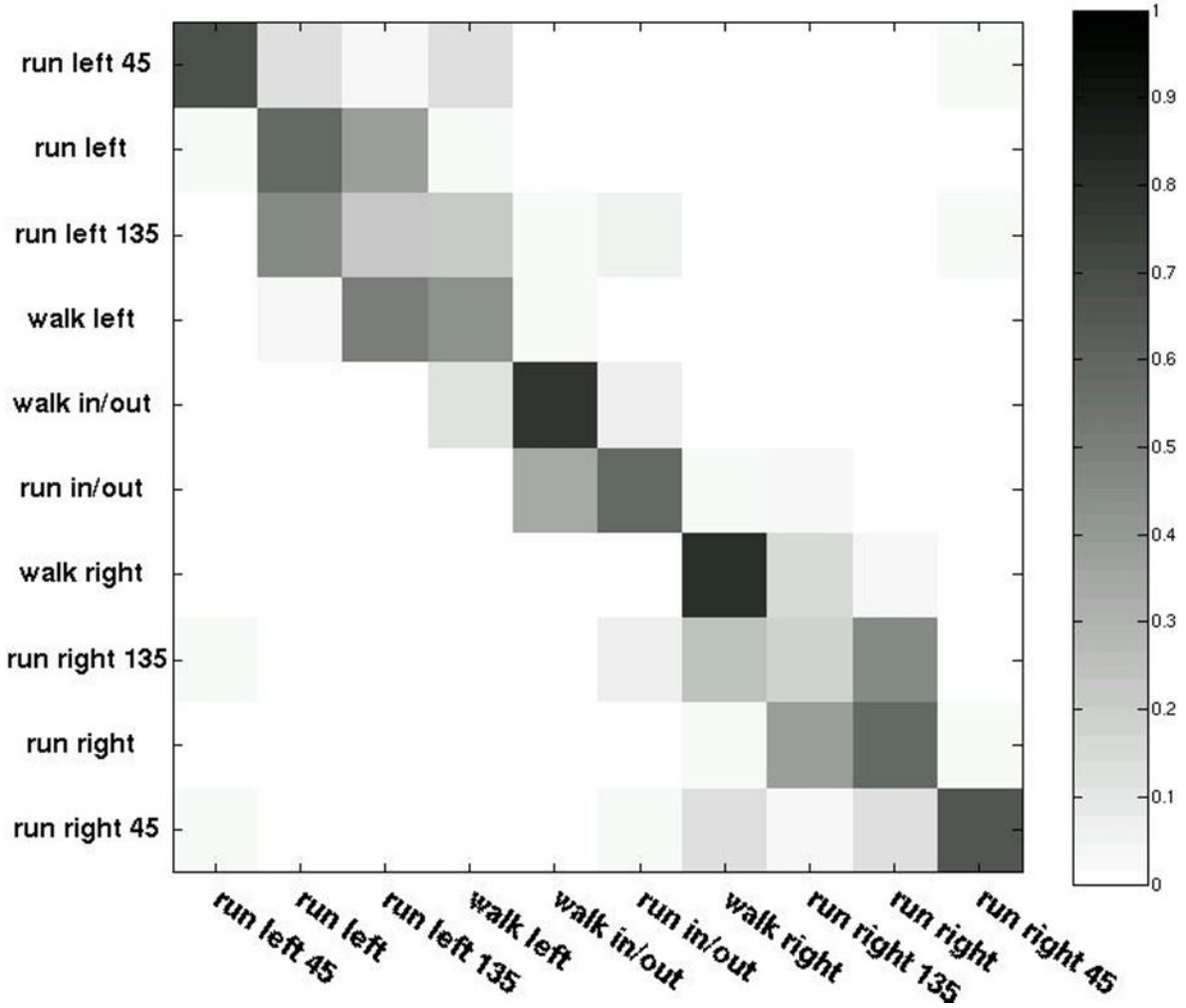
Исходное видео

Сопоставленные кадры



вход сопоставление

Классификация действий: футбол

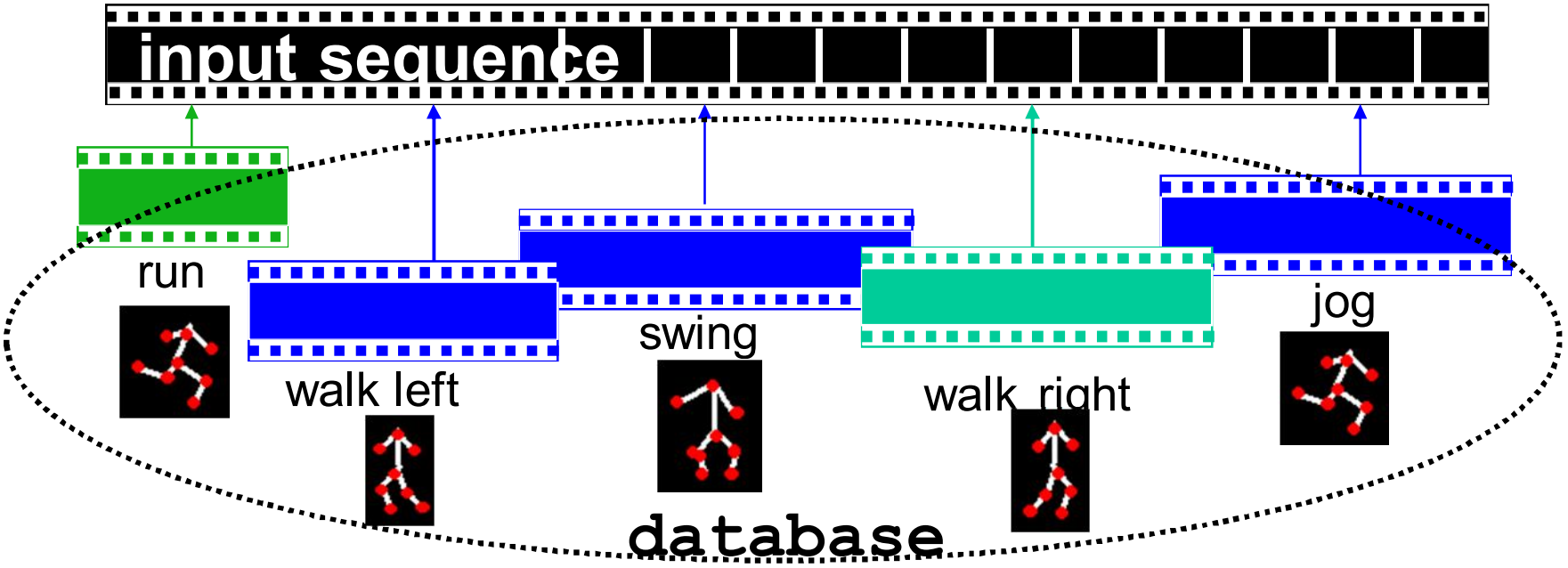


10 действий; 4500 кадров всего; 13 кадров в дескрипторе

Распознавание: теннис

Microsoft Research

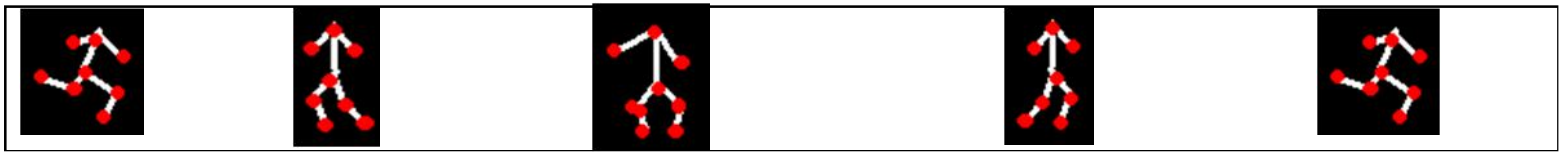
Аннотированная база событий



Распознавание действий:

run	walk left	swing	walk right	jog

Положения суставов:



Перенос 2D скелета

- База размечена 2D точками в местах сочленений конечностей (суставах)
- После сопоставления перенос скелета на новую последовательность
  - Уточнение (Поиск наилучшего сдвига/масштаба для совмещения карты движения)

Исходная последовательность:

Перенос скелета:

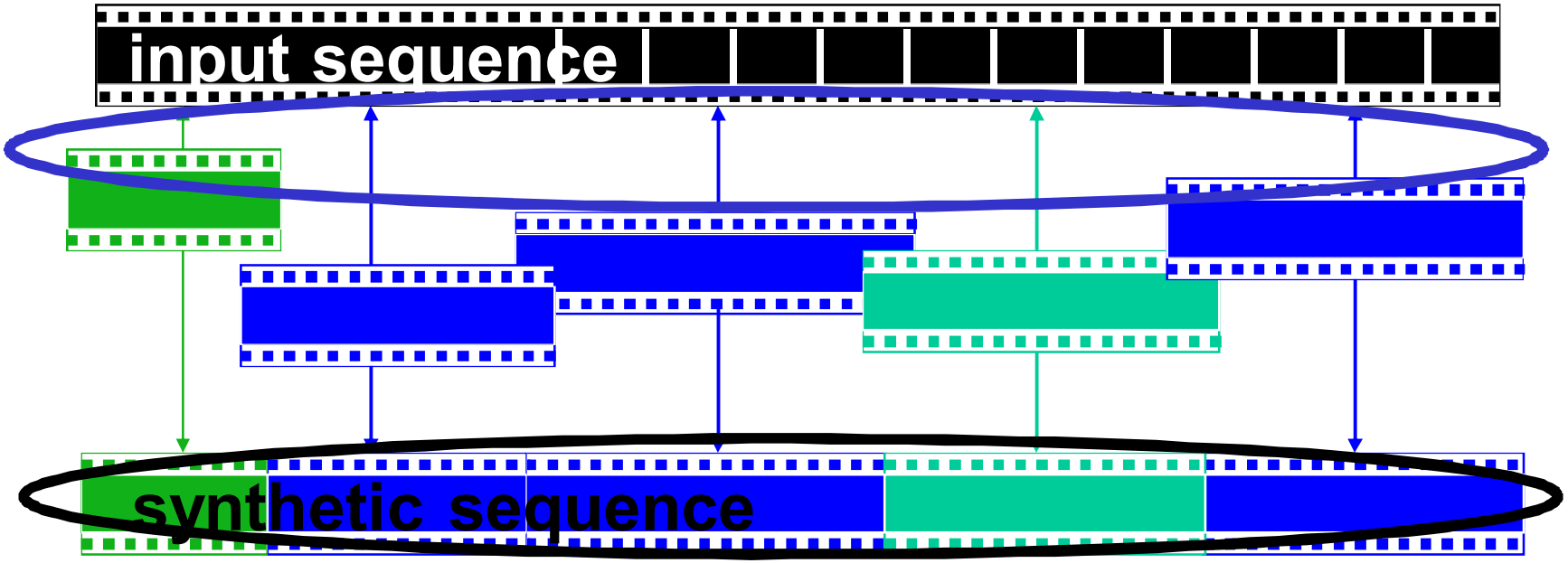
Перенос 3D скелета

- Дополняем базу фрагментов наборов движений фигур из палочек (stick-figures) из MoCap
- После сопоставления, получаем псевдо-3д модель движения!

Исходная последовательность:

Перенос скелета:

“Do as I Do” синтез движения



- Параметры качества синтеза:
  - Сходство движений между видео
  - Сходство внешности между видео
- Энергетическая формулировка, динамическое программирование

“Do as I Do”

Исходное движение

Результат

Исходный вид

3400 кадров

Замена персонажа

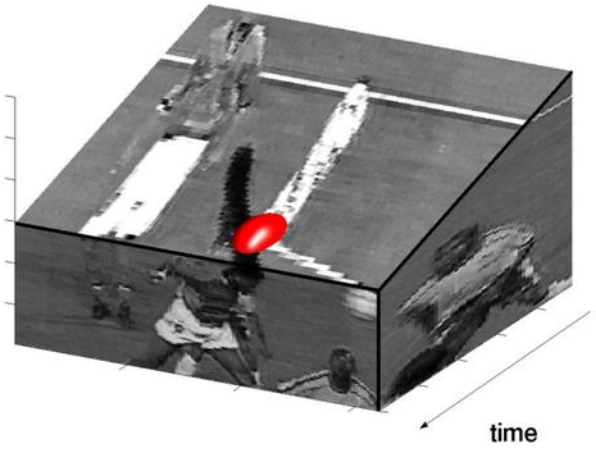
Microsoft Research

Особенности

- Применим формализм особых точек на изображении к видео
  - Harris3D (2003)
  - Cuboid (2005)
  - Hessian (2008)

H. Wang, M. M. Ullah, A. Kläser, I. Laptev and C. Schmid; "Evaluation of local spatio-temporal features for action recognition" *in Proc. BMVC'2009*,

Harris3D



Ivan Laptev, INRIA

I. Laptev and T. Lindeberg; "Space-Time Interest Points", *ICCV'2003*

Harris3D

- Вычислим матрицу вторых моментов, аналогично детектору Harris:

$$\mu(\cdot;\sigma,\tau) = g(\cdot;s\sigma,s\tau) * (\nabla L(\cdot;\sigma,\tau)(\nabla L(\cdot;\sigma,\tau))^T)$$

Где g – ядро гаусса (сглаживание)
σ, τ - параметры масштаба

$$\mu = g(\cdot;\sigma_i^2,\tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix},$$

- Функция отклика угла:

$$H = \det(\mu) - k\,\mathrm{trace}^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3.$$

Microsoft Research

Выбор масштаба

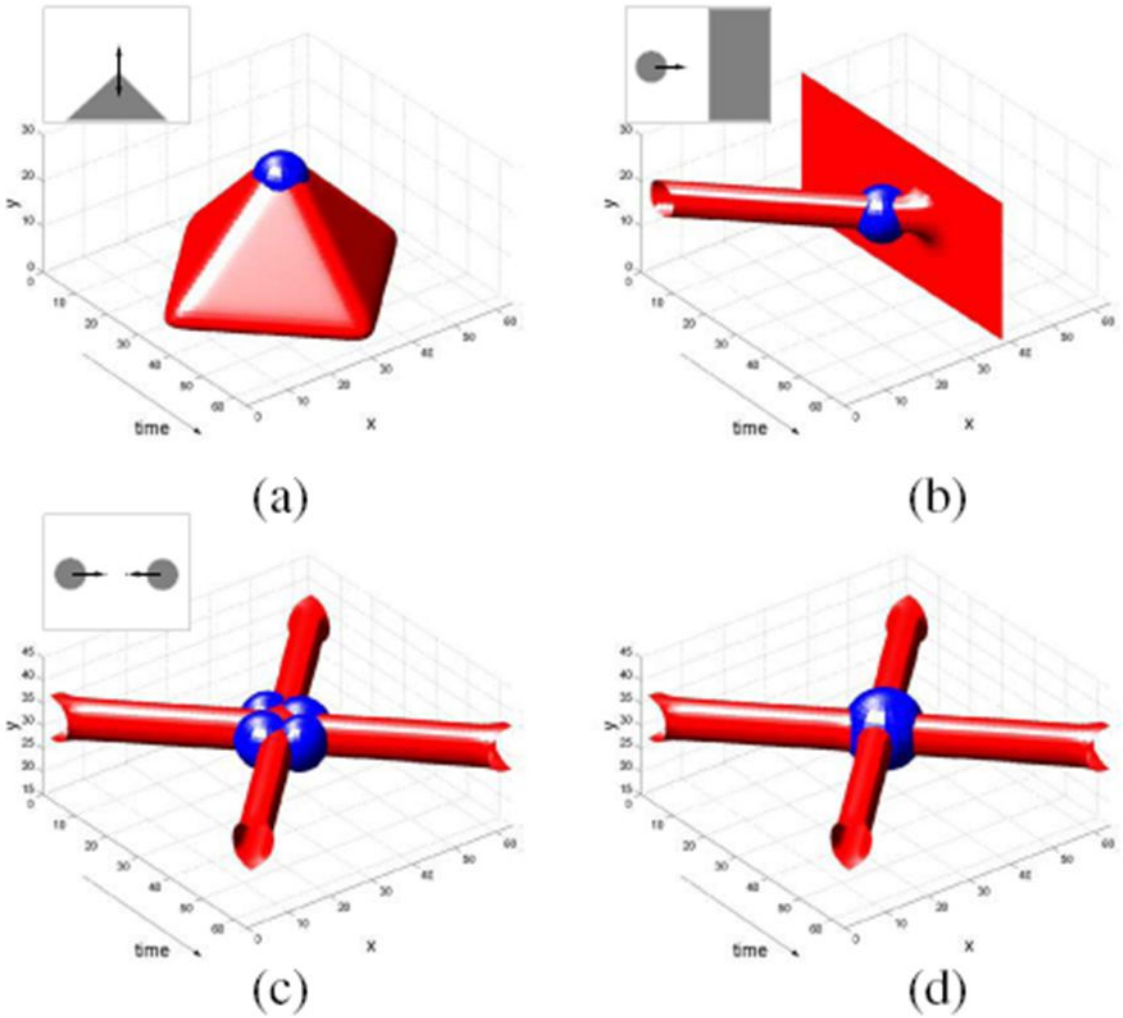
Лапласиан:

$$\nabla^2_{norm} L = L_{xx,norm} + L_{yy,norm} + L_{tt,norm},$$

Нормализованный вариант:

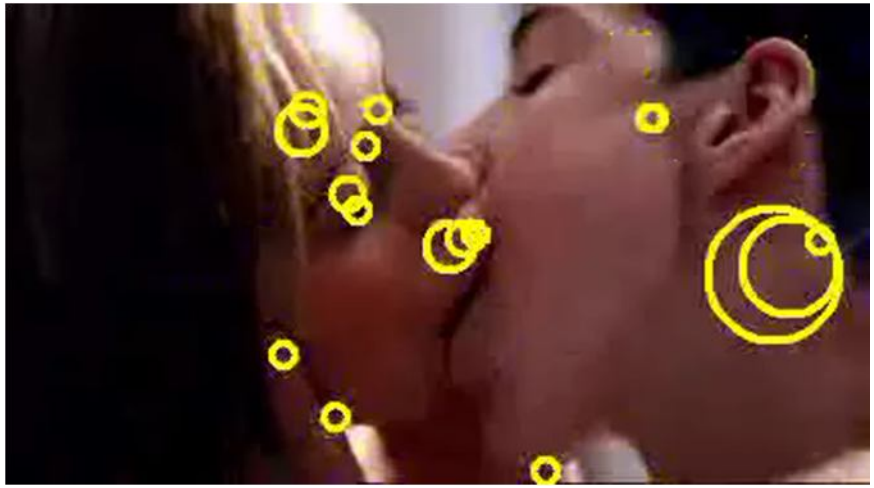
$$\nabla^2_{norm} L = \sigma^2 \tau^{1/2} (L_{xx} + L_{yy}) + \sigma \tau^{3/2} L_{tt}.$$

- Как при поиске инвариантных точек на изображениях
  - Ищем локальные максимумы на фиксированных масштабах
  - Уточняем масштаб
  - Уточняем положение в пространстве



Синтетические примеры

Примеры для кино



Cuboid

- Cuboid
  - Функция отклика угла:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

  - Где g – фильтр Гаусса по изображению
    h – фильтр Габора

$$h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2} \qquad \omega = 4/\tau$$

$$h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$$

- Hessian
  - Вычисление 3D матрицы гессиана

• P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In VS-PETS, 2005.
• G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In ECCV, 2008.

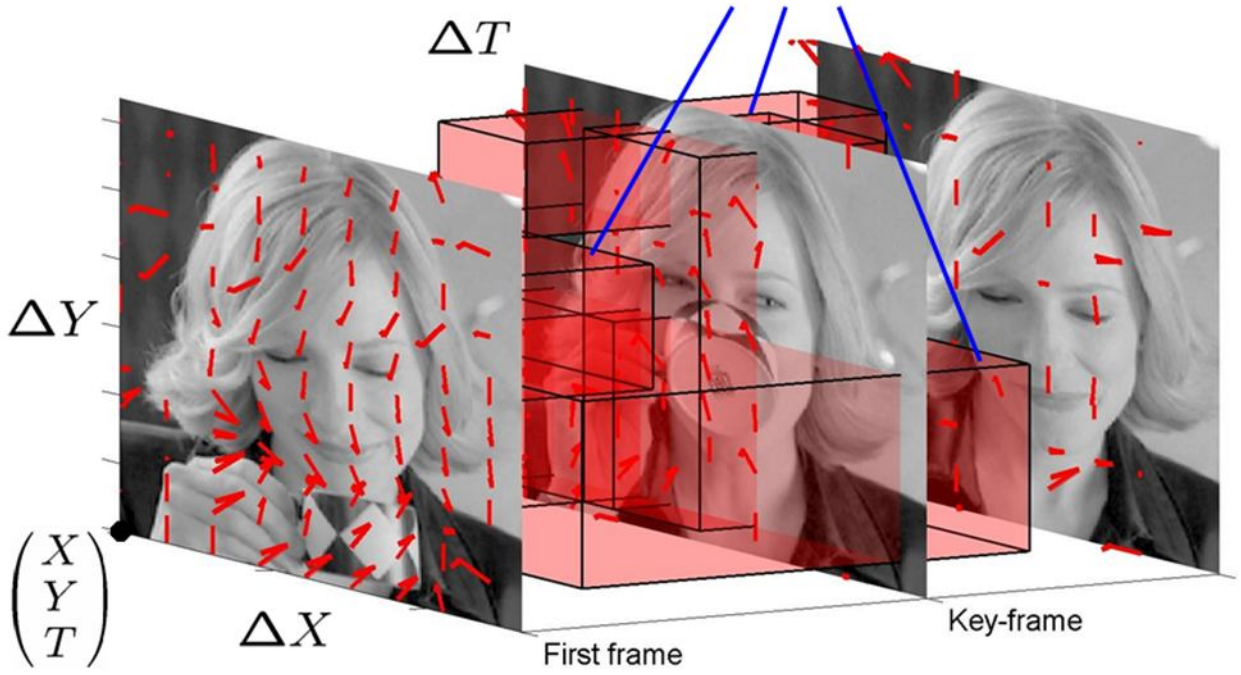
Код

- Ivan Laptev’s code
  - http://www.irisa.fr/vista/Equipe/People/Laptev/download.html#stip
- Piotr's Image & Video Toolbox for Matlab
  - http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html
  - Много полезных функций (k-means, meanshift, PCA, ferns, RBF, DOG-фильтры и т.д.)
- Hessian executables:
  - http://homes.esat.kuleuven.be/~gwillems/research/Hes-STIP/

Microsoft Research

Плотный выбор точек

- По аналогии с обычными изображениями, можно точки выбирать плотно на изображении, а не искать специальным детектором
- Обычно выбирают с 50% перекрытием



Microsoft Research

Дескрипторы

- Jet (2003)
  - Вектор градиентов:

$$j = (L_x, L_y, L_t, L_{xx}, ..., L_{ttt}).$$

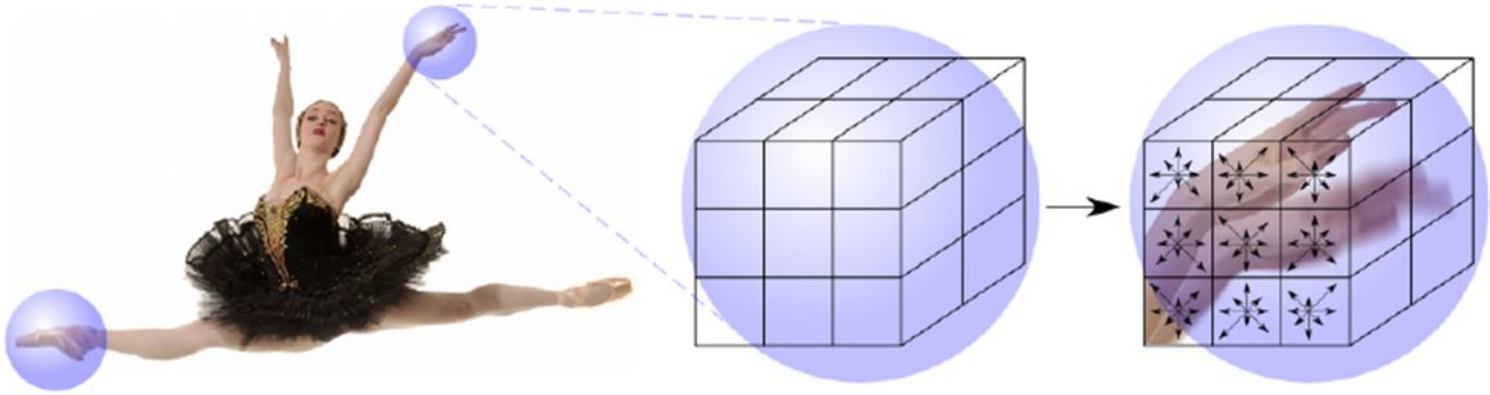
  - Расстояние Махаланобиса

$$d^2(j_1, j_2) = (j_1 - j_2)\Sigma^{-1}(j_1 - j_2)^T$$

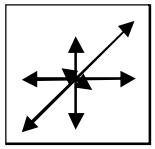
- Cuboid (2005)
- HOG/HOF (2007)
- HOG3D (2008)
- Extended SURF (2008)

HOG/HOF

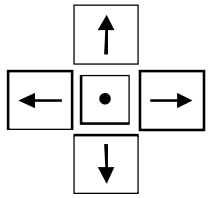
Патчи по окрестностям



Histogram of oriented spatial grad. (HOG)



Histogram of optical flow (HOF)



3x3x2x4bins **HOG** descriptor

3x3x2x5bins **HOF** descriptor

Public code available at www.irisa.fr/vista/actions

Microsoft Research

Дескрипторы

- Cuboid
  - Вычисление градиентов для каждого пикселя во фрагменте
  - PCA -> 100 первых компонент для дескриптора

- HOG3D
  - 3D аналог для HOG
  - Вычисляем градиенты в каждом пикселе объёма
  - Дискретизируем, сопоставляя сторонам правильного многоугольника
  - Разбиваем весь объем на ячейки
  - В каждой ячейке считаем гистограмму направлений

Microsoft Research

Тестовые базы

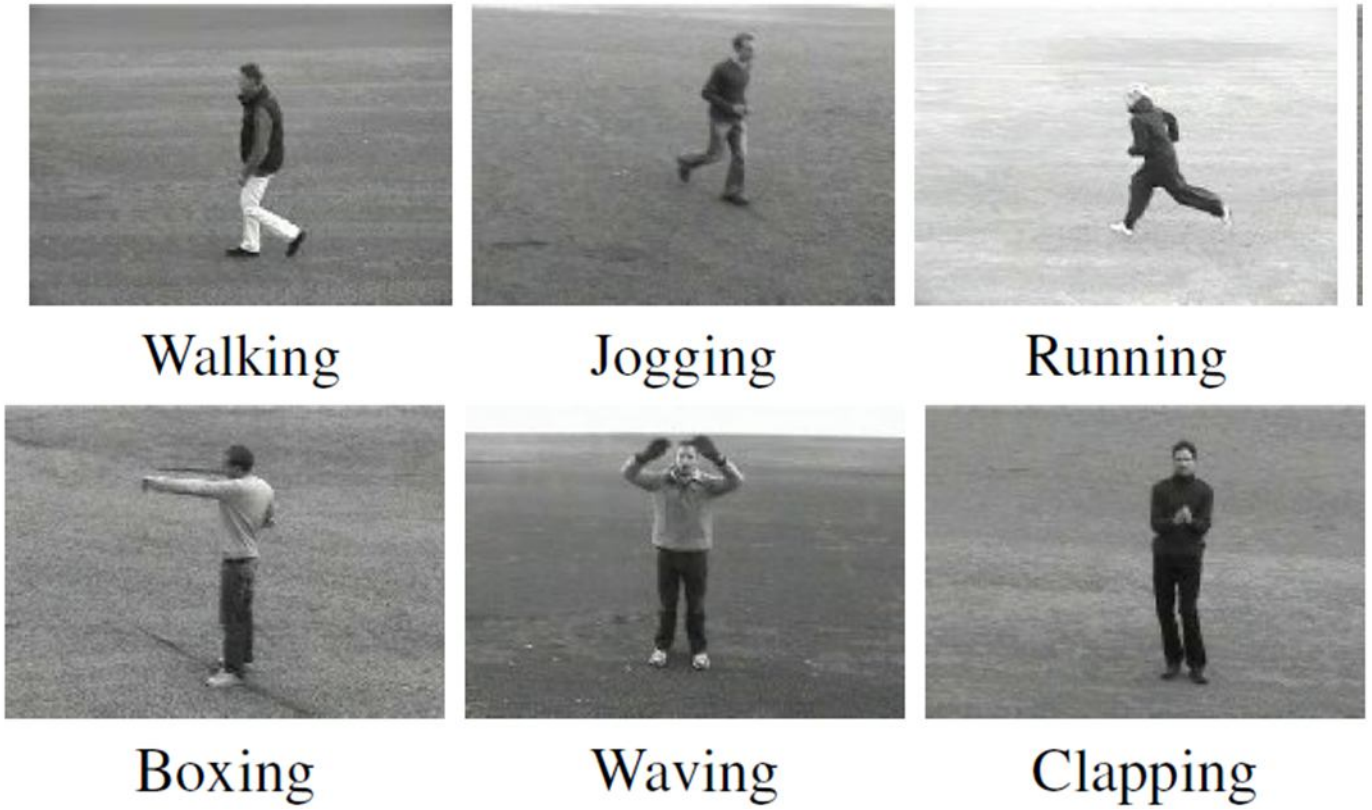
- Самые известные:
  - KTH Actions
  - UCF Sport Actions
  - Hollywood2

• C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In ICPR, 2004.

•M. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In CVPR, 2008

•M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In CVPR, 2009

KTH Actions



- 25 человек, 6 действий, 4 сценария (внутри, снаружи, снаружи др. масштаба, снаружи в др. одежде)
- Всего 2391 фрагмент

UCF Sport Actions


Diving


Kicking


Walking


Skateboarding


High-Bar-Swinging

- 10 видов «спортивных» действий
- 150 видеофрагментов
- Можно увеличить с помощью «шевеления»/зеркального отображения

Hollywood2

AnswerPhone GetOutCar HandShake HugPerson Kiss

- 10 разных обыденных действий из 69 голливудских фильмов
- 1707 фрагментов

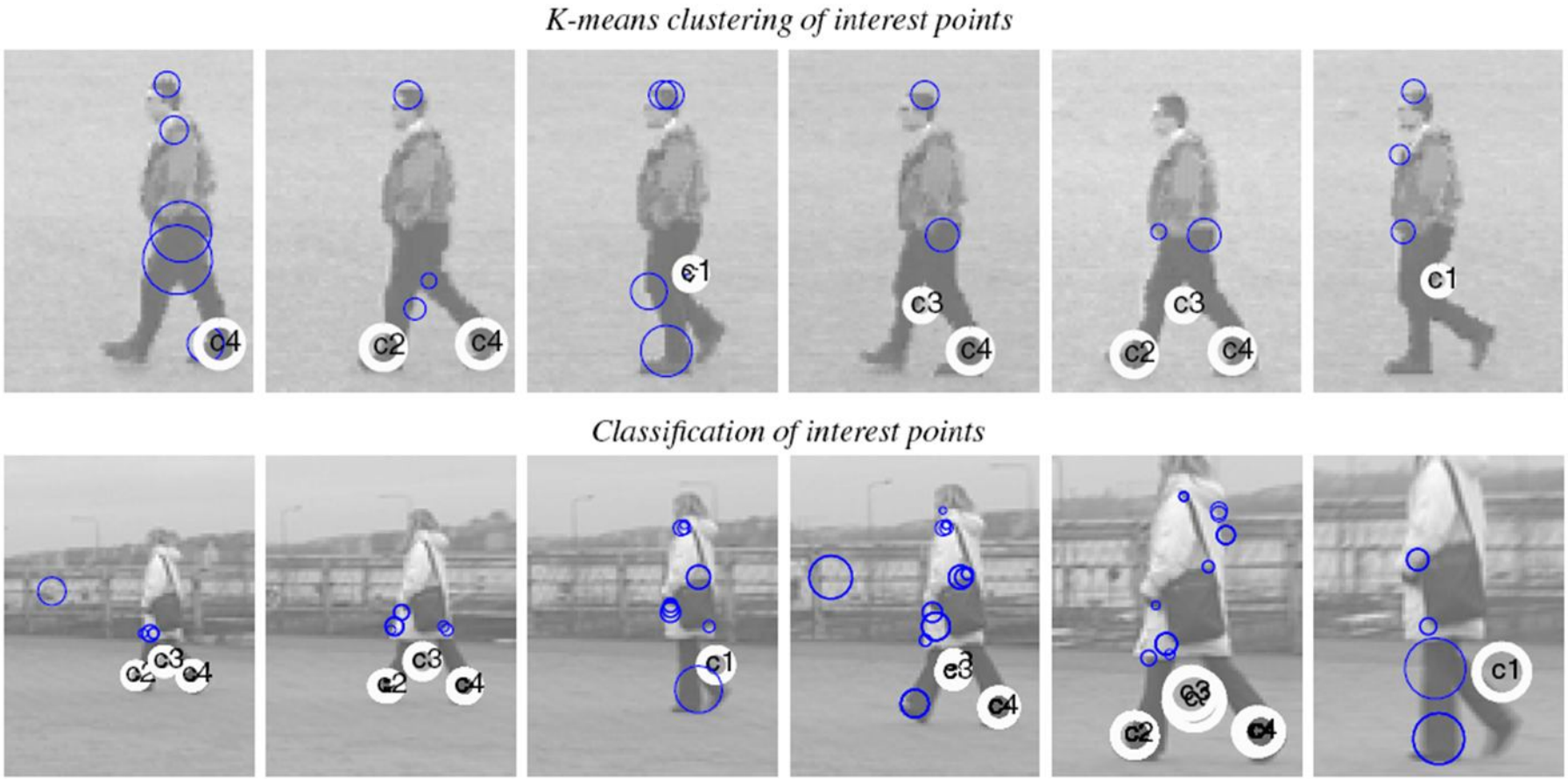
Эксперименты со STIP

- Построение мешка слов для STIP
  - Перенос силуэтов
- Классификация по мешку слов
  - SVM с ядром Хи-Квадрат

$$K(H_i, H_j) = exp(-\frac{1}{2A}\sum_{n=1}^{V}\frac{(h_{in}-h_{jn})^2}{h_{in}+h_{jn}})$$

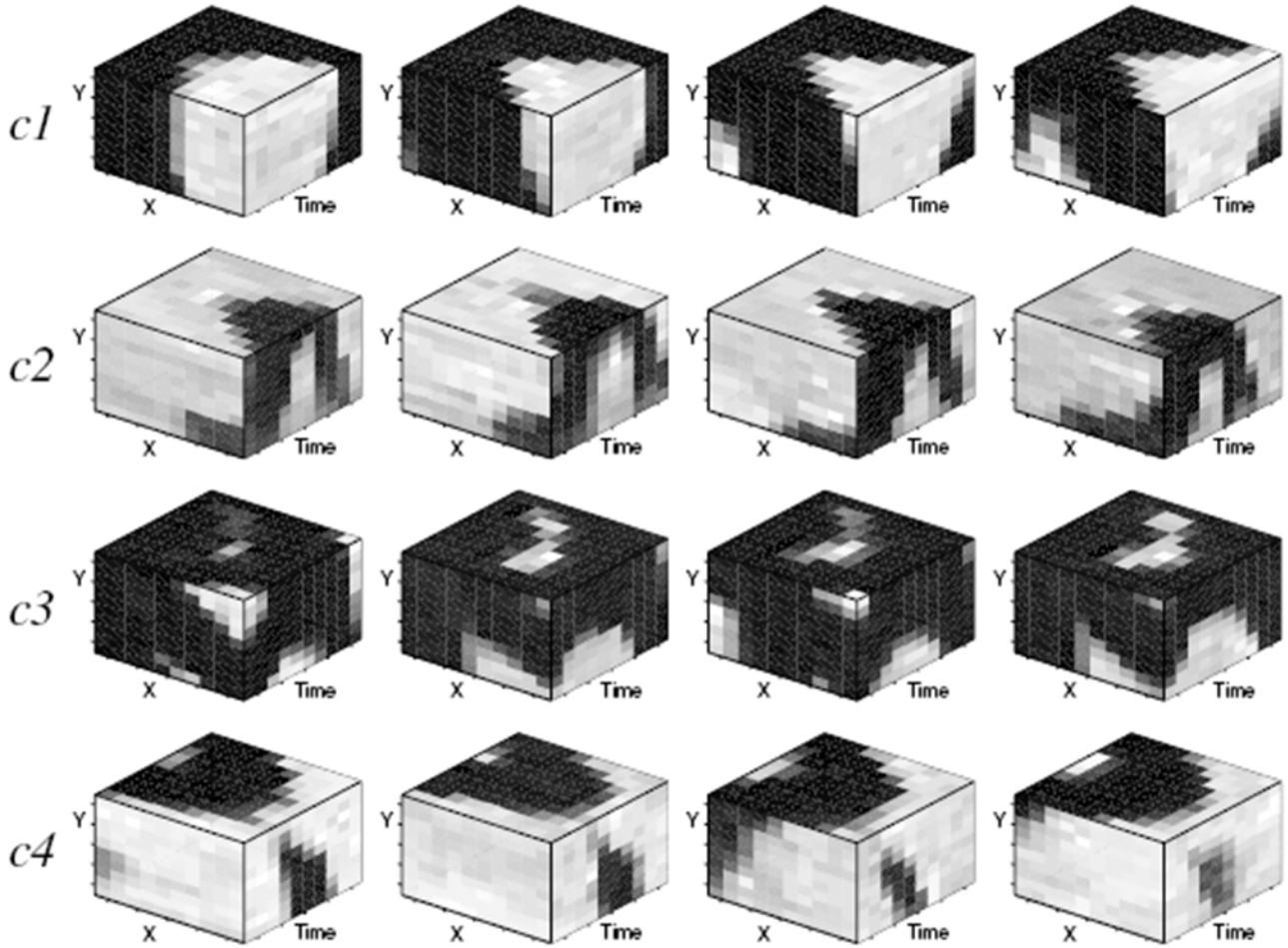
  - Сравнение разных детекторов / дескрипторов

Классификация интересных точек



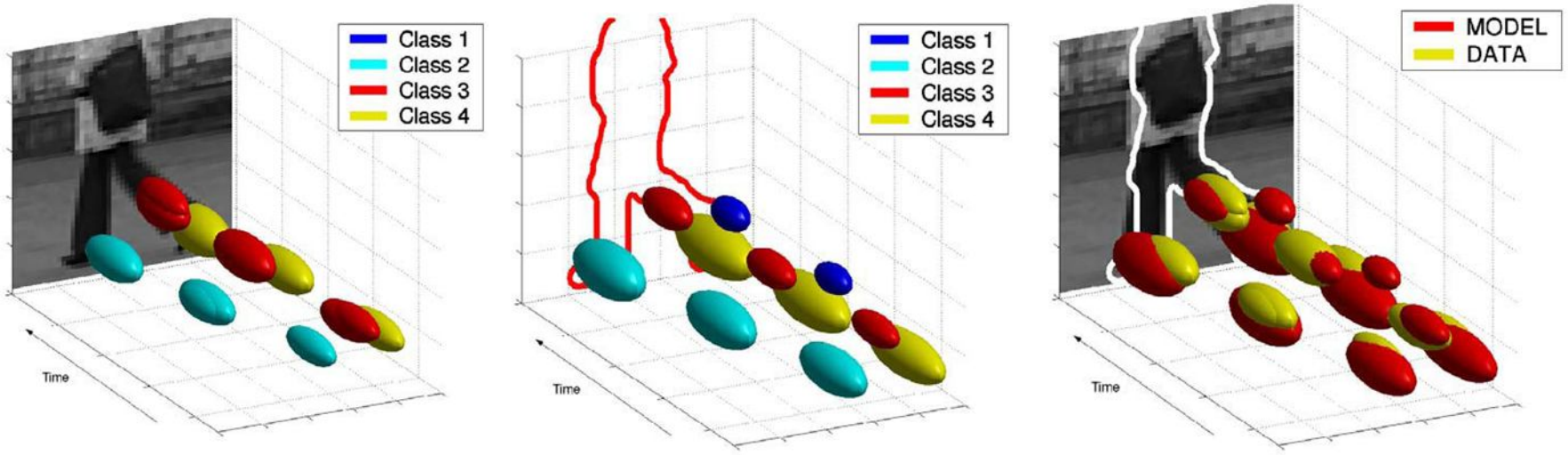
- Кластеризуем К- средними
- Квантуем особенности

Примеры



- Примеры 4х самых часто встречающихся пространственно-временных особенностей

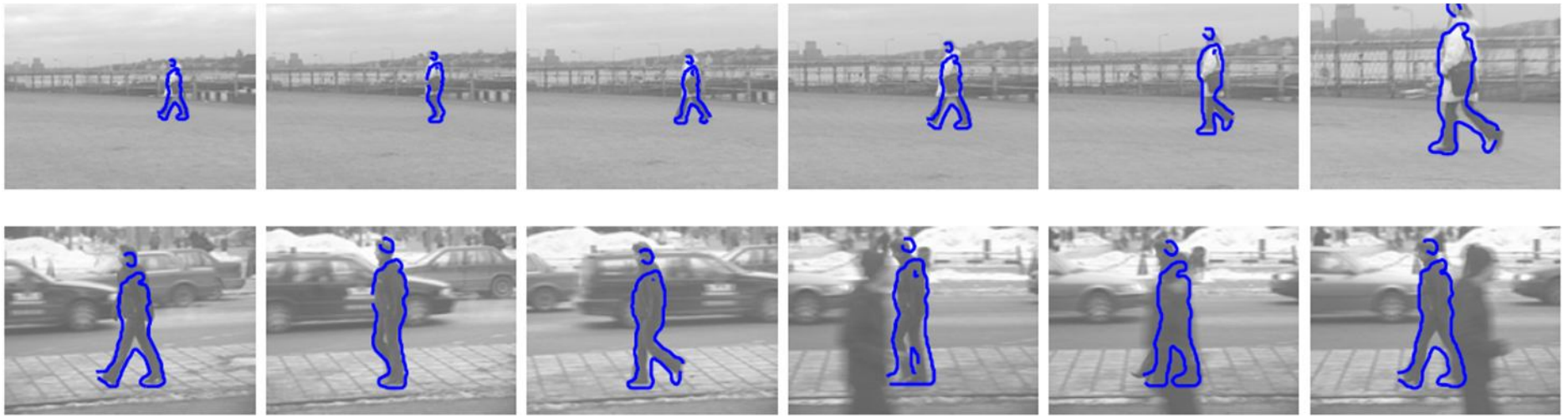
Перенос силуэтов



- Выделяем силуэты в тренировочных данных
- Выделяем STIP в видеофрагменте
- Находим ближайший набор STIP в базе
- Совмещаем и переносим силуэт

Пример

Microsoft Research



Сравнение

Microsoft Research

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	89.0%	**91.8%**	80.9%	**92.1%**	–	–
Cuboids	**90.0%**	88.7%	82.3%	88.2%	89.1%	–
Hessian	84.6%	88.7%	77.7%	88.6%	–	81.4%
Dense	85.3%	86.1%	79.0%	88.0%	–	–

База KTH Actions

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	79.7%	78.1%	71.4%	75.4%	–	–
Cuboids	**82.9%**	77.7%	72.7%	76.7%	76.6%	–
Hessian	79.0%	79.3%	66.0%	75.3%	–	77.3%
Dense	**85.6%**	81.6%	77.4%	**82.6%**	–	–

База UCF Sports Actions

Сравнение

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	43.7%	45.2%	32.8%	43.3%	–	–
Cuboids	45.7%	**46.2%**	39.4%	42.9%	45.0%	–
Hessian	41.3%	**46.0%**	36.2%	43.0%	–	38.2%
Dense	45.3%	**47.4%**	39.4%	45.5%	–	–

База Hollywood2

	HOG3D	HOG/HOF	HOG	HOF
reference	43.7%	45.2%	32.8%	43.3%
w/o shot boundary features	43.6%	**45.7%**	35.3%	43.4%
full resolution videos	45.8%	**47.6%**	39.7%	43.9%

База Hollywood2 – разное разрешение

Сравнение

Microsoft Research

Spatial Size	Hollywood2				UCF			
	HOG3D	HOG/HOF	HOG	HOF	HOG3D	HOG/HOF	HOG	HOF
18×18	**45.3%**	47.4%	39.4%	45.5%	**85.6%**	**81.6%**	**77.4%**	82.6%
24×24	45.1%	**47.7%**	**39.4%**	**45.8%**	82.0%	81.4%	76.8%	**84.0%**
36×36	44.8%	47.3%	36.8%	45.6%	78.6%	79.1%	76.5%	82.4%
48×48	42.8%	46.5%	35.8%	45.5%	78.8%	78.6%	73.9%	79.0%
72×72	39.7%	45.2%	32.2%	43.0%	77.8%	78.8%	69.6%	78.4%

Размеры окрестностей

	Harris3D + HOG/HOF	Hessian + ESURF	Cuboid Detector + Descriptor	Dense + HOG3D	Dense + HOG/HOF
Frames/second	1.6	4.6	0.9	0.8	1.2
Features/frame	31	19	44	643	643

Оценка скорости

Распознавание действий в кино

I. Laptev and P. Pérez. "Retrieving actions in movies“ *ICCV 2007*

Проблемы

Пить

Курить

Разница в форме

Разница в движении

Оба действия похожи по форме (поза человека) и по движению (движение руки)

Вариабельность больше, чем для изображений

Но движение дает дополнительную информацию

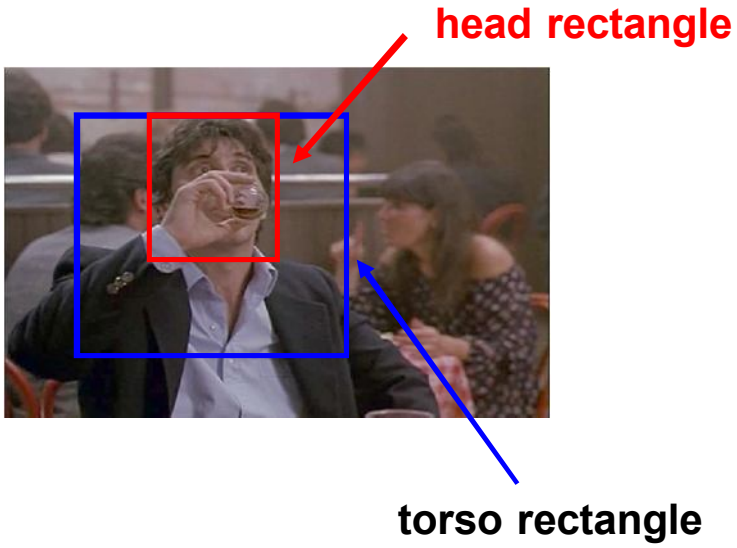
Тестовая база

- В этой работе впервые была сделана попытка распознавать действия в фильмах: “Coffee and Cigarettes”; “Sea of Love”
- Первая размеченная база действий из кинофильмов

“*Пить*”: 159 размеченных примеров

“*Курить*”: 149 размеченных примеров

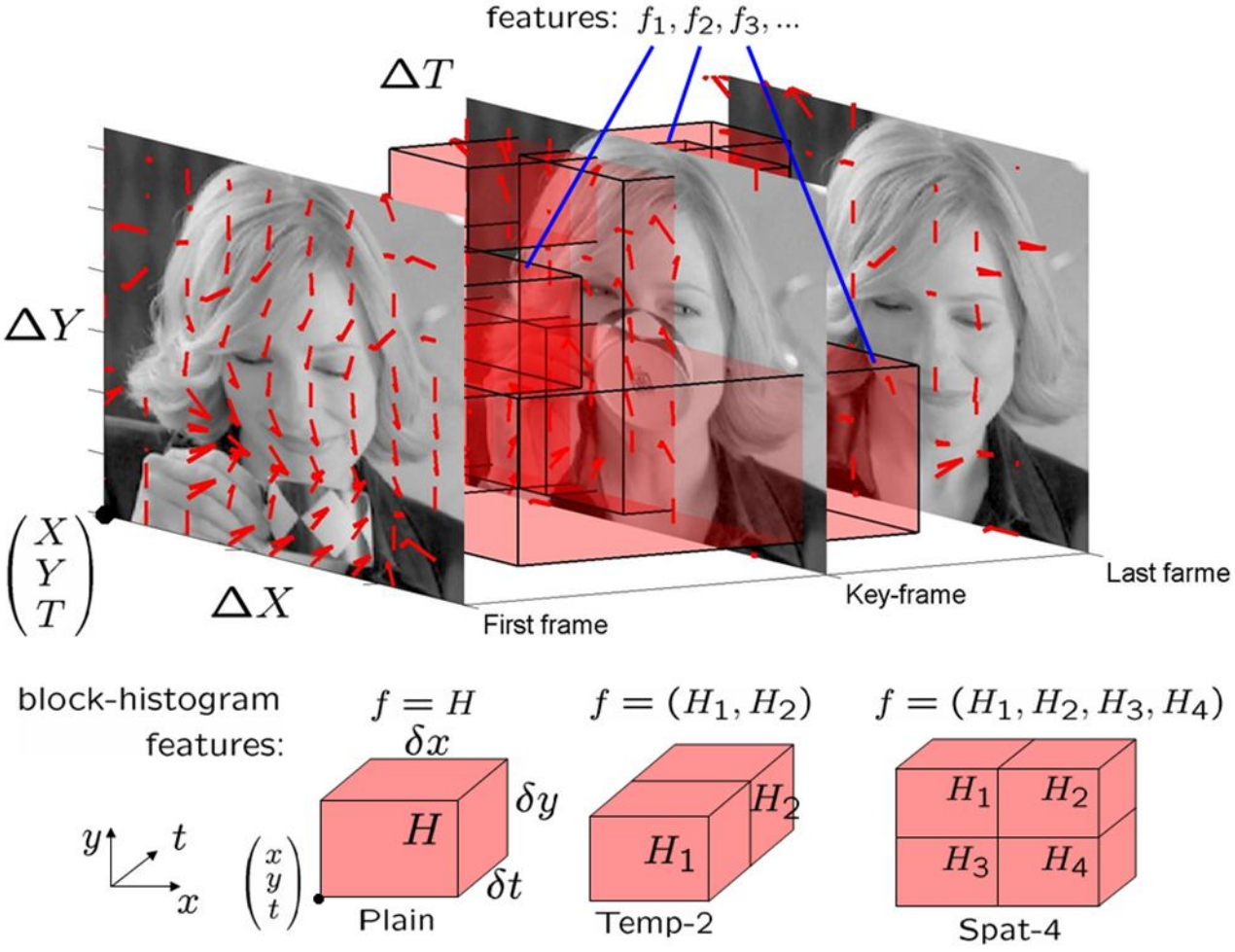
Пространственная разметка



Временная разметка

Примеры действия «пить»

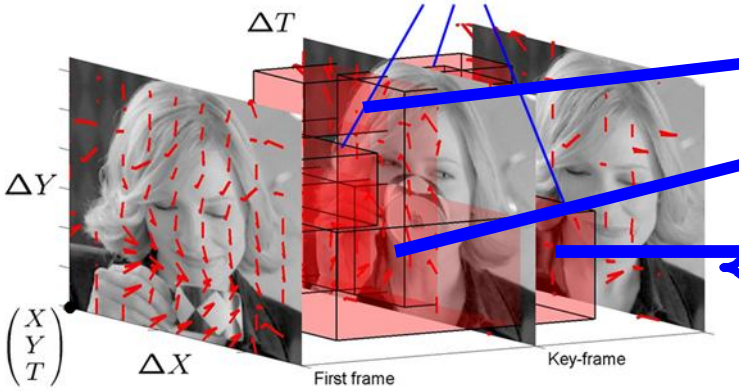
Признаки



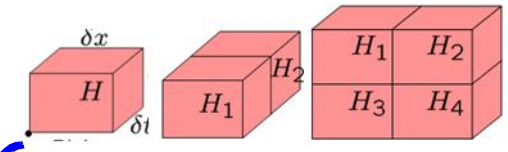
- Весь объем разбиваем на ячейки, всего 14x14x8
- Каждую ячейку нормализуем до 5x5x5 пикселей
- На ячейках можно определить большой набор признаков (features)

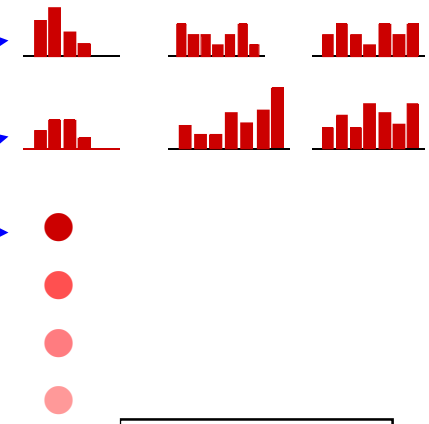
- Features – (x,y,t, dx, dy, dt, β, ψ), где (x,y,t) – начало блока, dx,dy,dt – размер, β – вид блока, ψ – тип признака

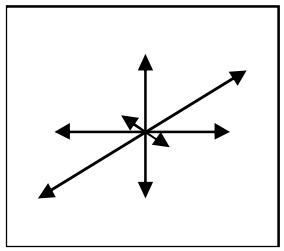
Признаки



HOG: histograms of **oriented gradient**

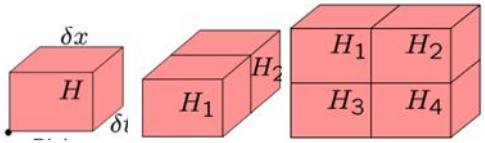


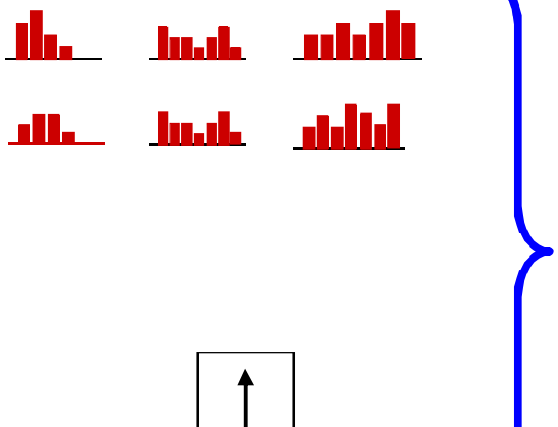


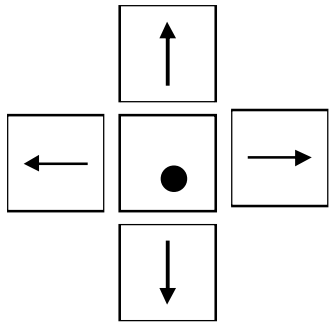


(grad4) 4 grad. orientation bins

HOF: histograms of **optic flow**







(OF5) 4 OF direction bins + 1 bin for no motion

Классификация

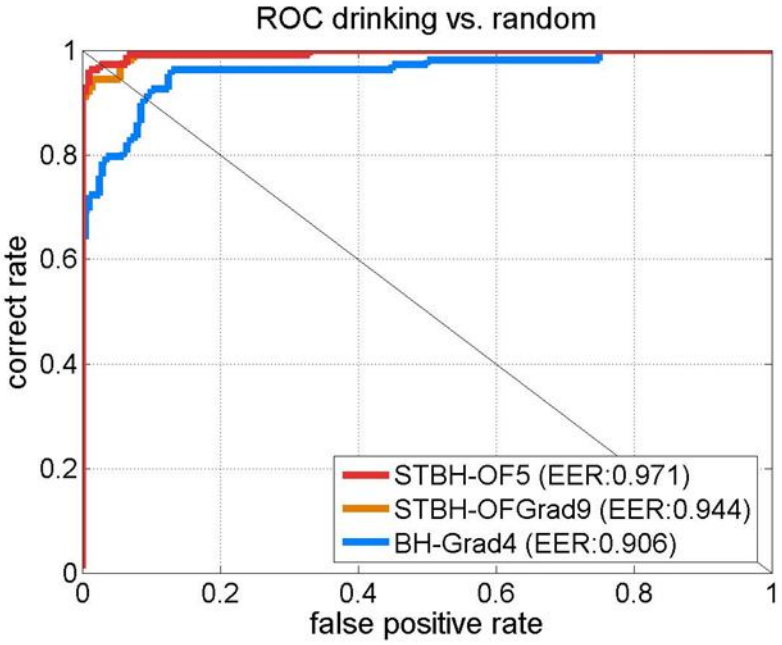
Microsoft Research

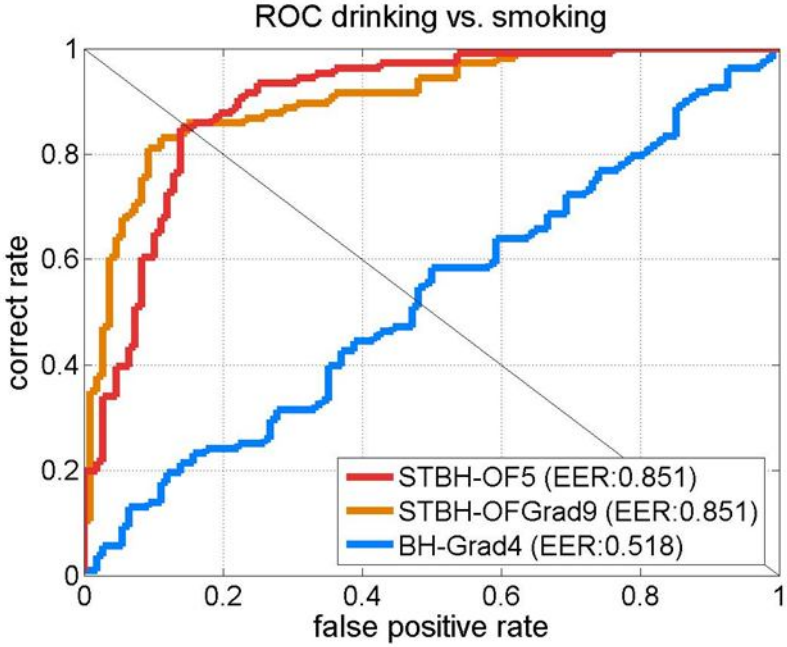
- Распознавание объёмов
  - Всего порядка ~10^7 разных признаков можно определить (очень много)
  - Используем AdaBoost
  - На каждом шаге выбираем 10^3 признаков случайным образом

- Можем аналогично распознавать действие по статичным кадрам (Keyframe detector)
  - Признаки HOG, AdaBoost

Результаты классификации действий

Random motion patterns



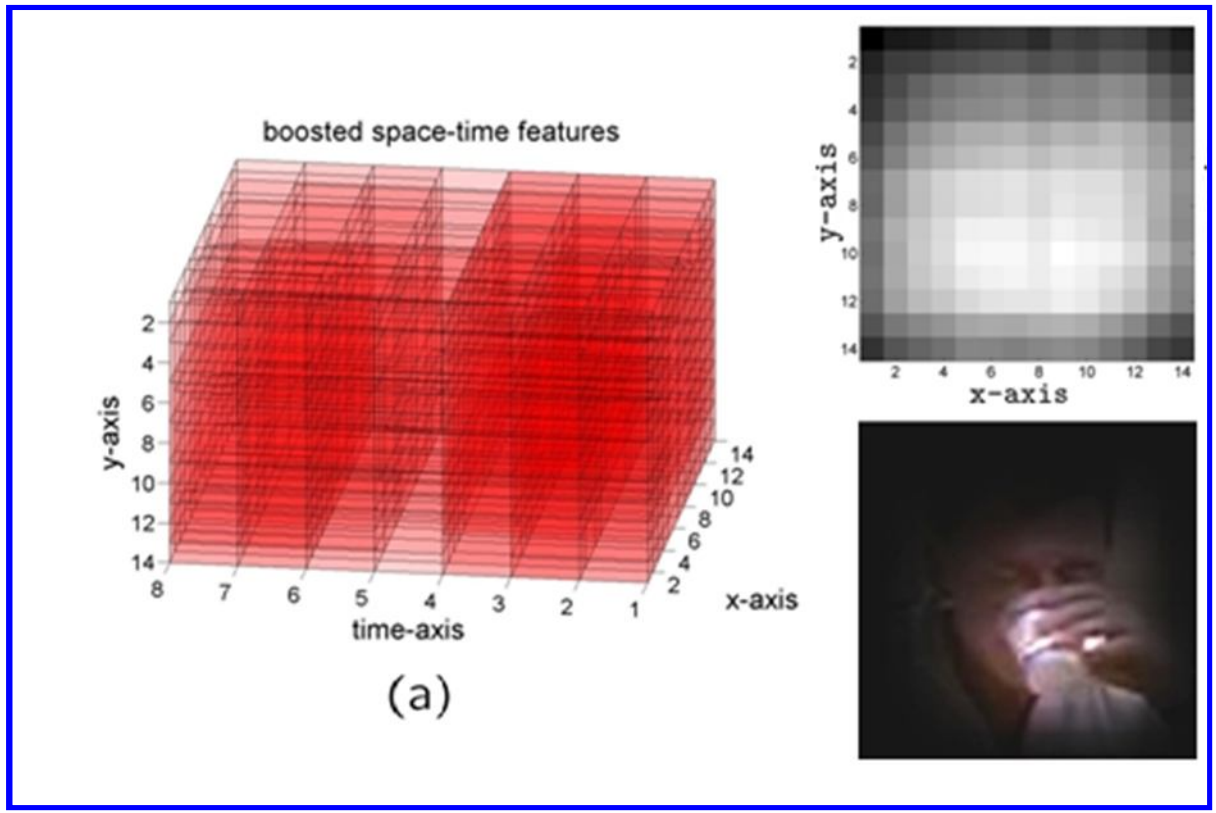


- Дополнительная информация о форме не улучшает работу пространственно-временного классификатора

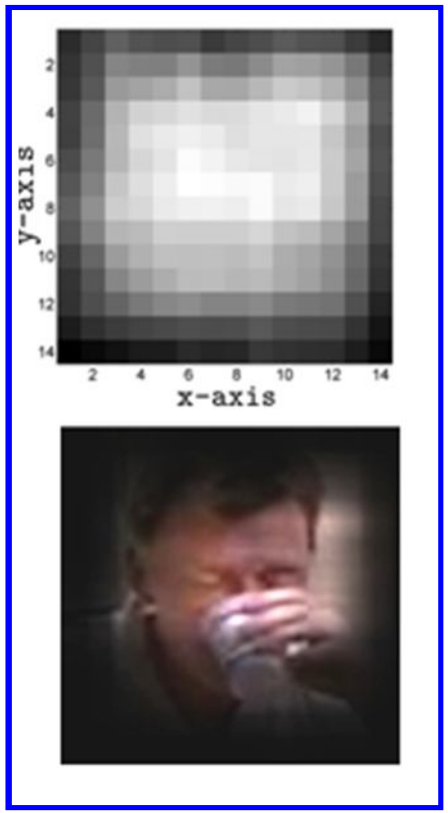
Свойства классификатора

- Сравниваем используемые особенности:
  - Space-time классификатор (HOF features)
  - Статичный классификатор ключевых кадров (HOG features)

Выход: суммарные карты признаков



Space-time classifier



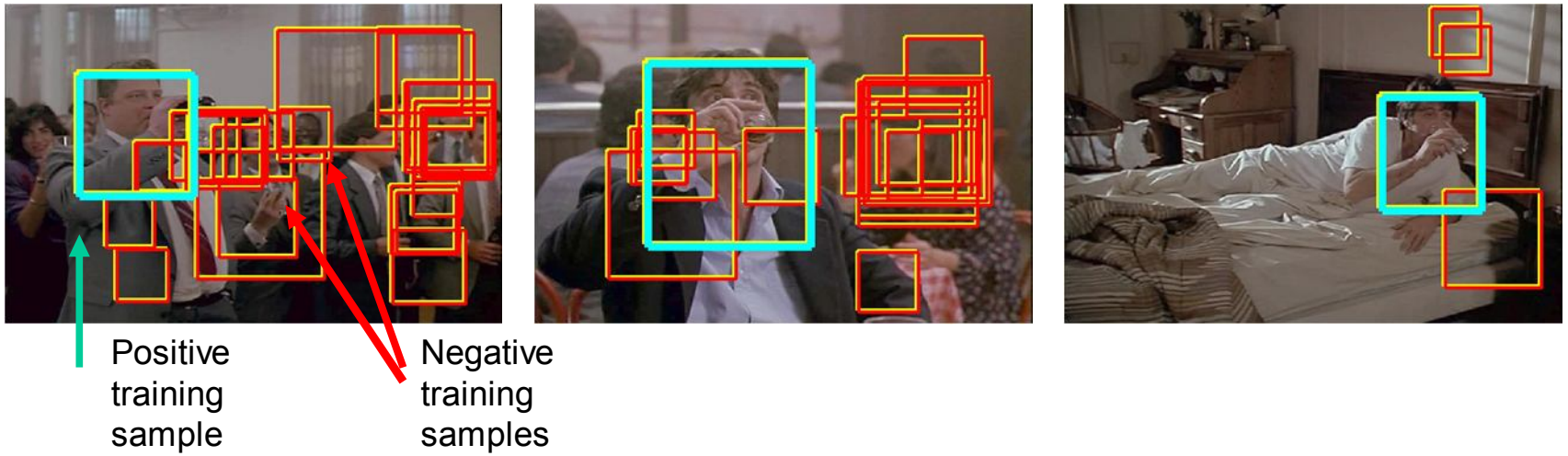
Static keyframe classifier

Сравнение со STIP

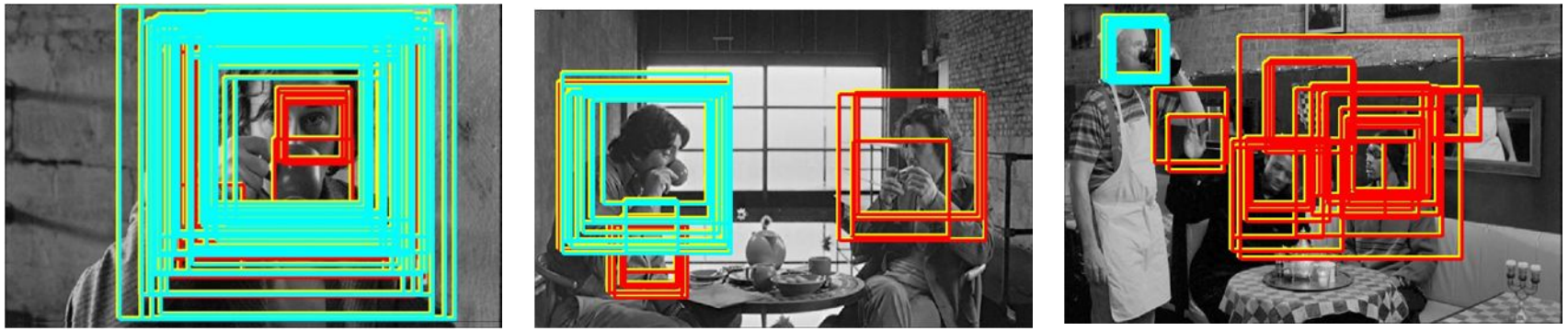
Keyframe priming

Идея: Будем распознавать только там, где нашел keyframe-детектор

Обучение



Тестирование



Поиск действий

Тестовый набор:

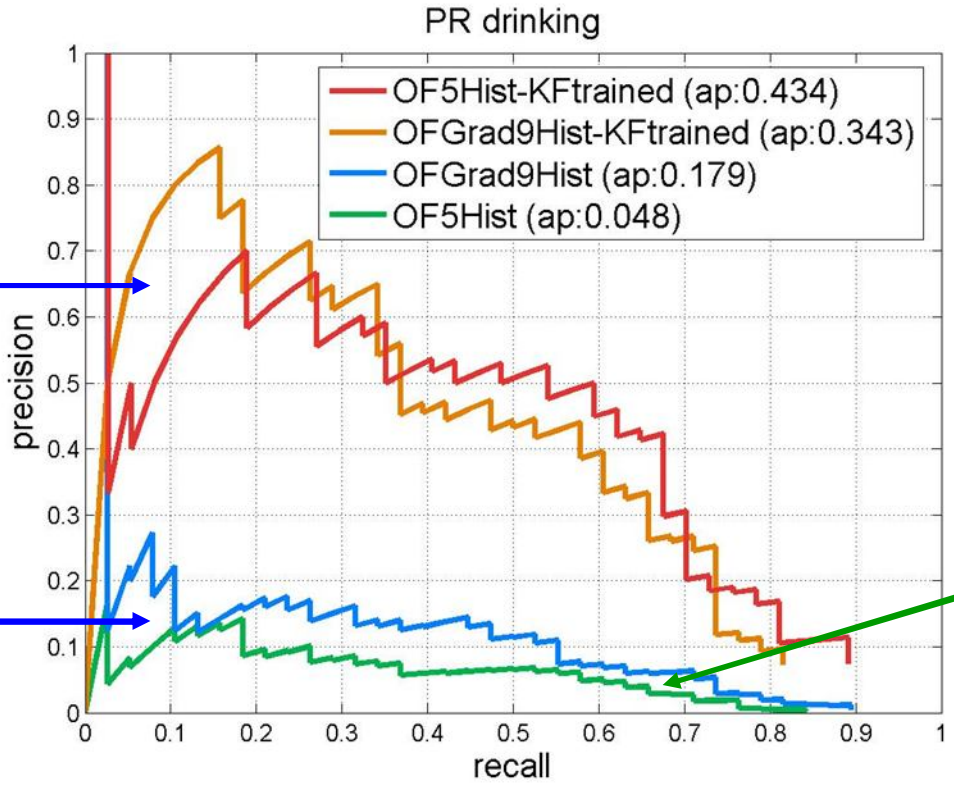
- 25мин из “Coffee and Cigarettes” с 38 действиями «пить»
- Нет перекрытия с обучающей выборкой по сценам и людям

Обнаружение:

- поиск по всевозможных пространственно-временным положениям и диапазонам (длинам фрагментов)

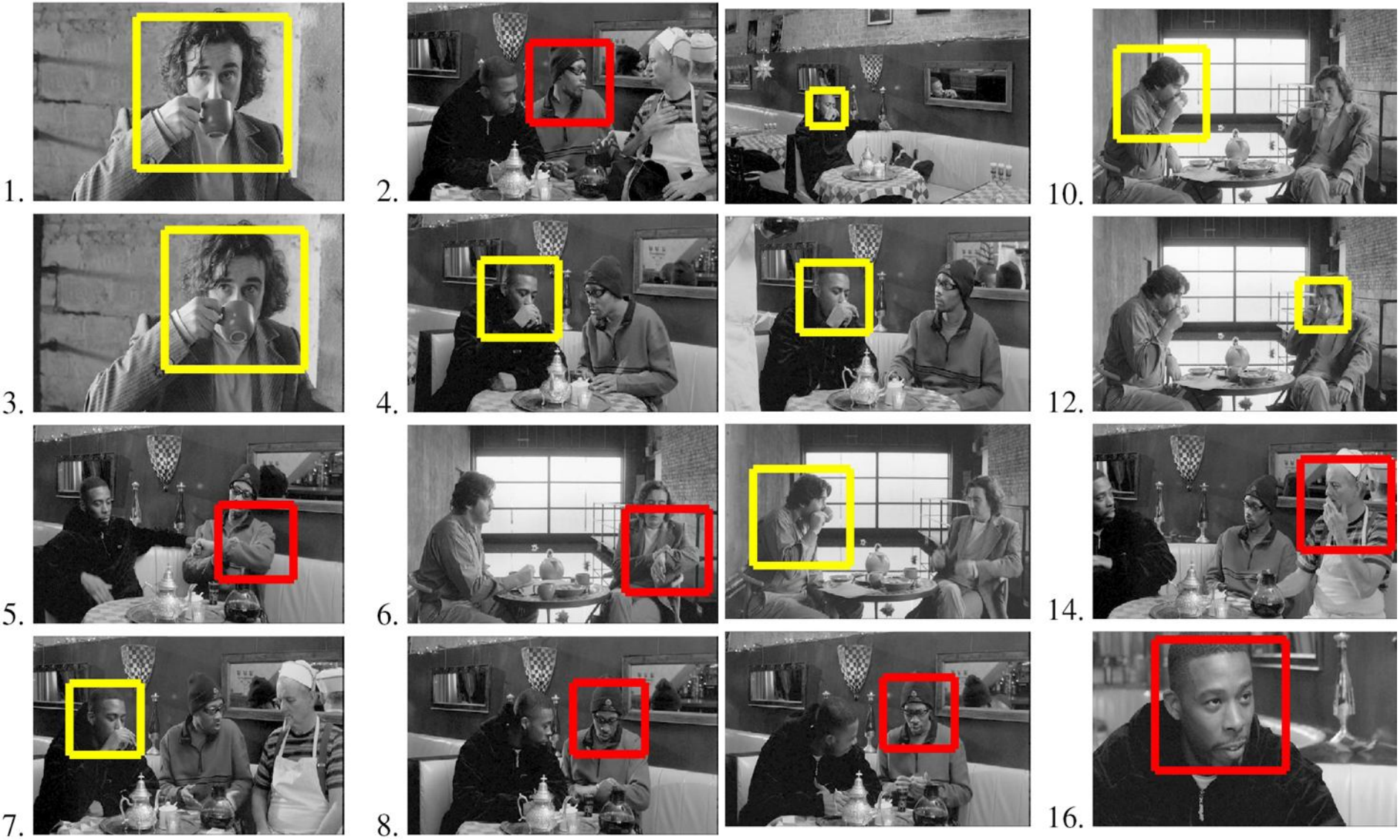
Keyframe priming

No Keyframe priming



Similar approach to Ke, Sukthankar and Hebert, ICCV05

Примеры работы

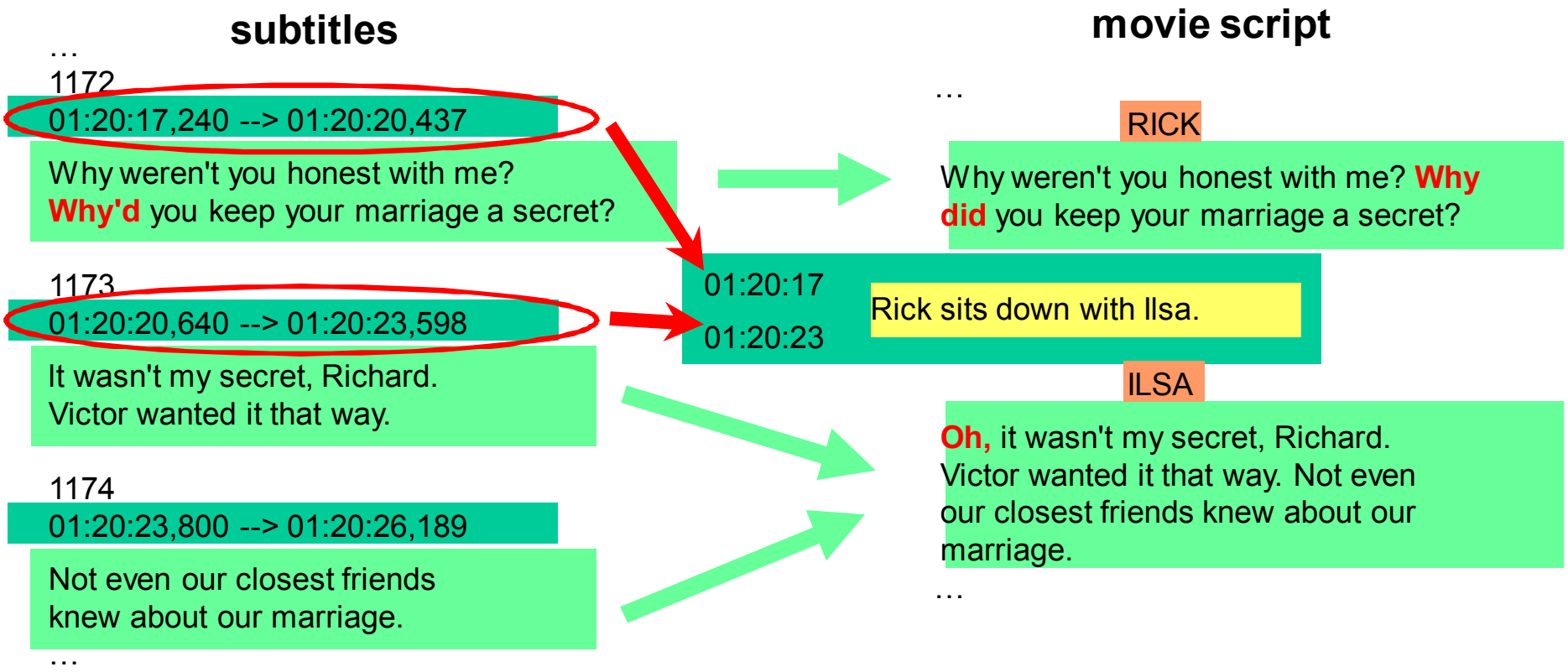


Распознавание действий в кино

I. Laptev, M. Marszałek, C. Schmid and B. Rozenfeld; "Learning realistic human actions from movies“ *CVPR 2008*

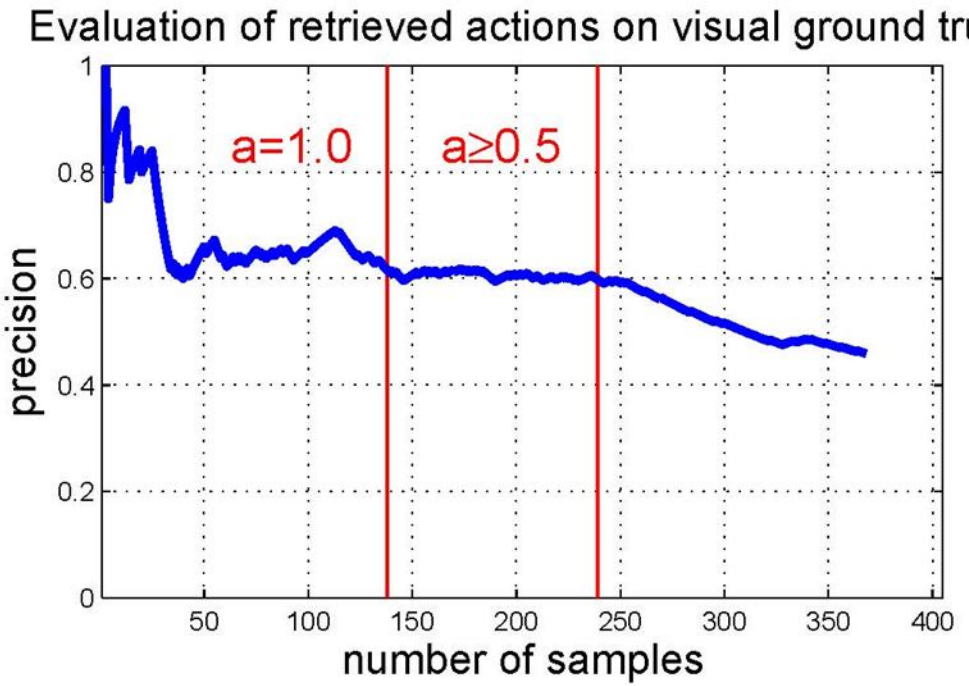
Аннотация по сценарию

- Сценарии есть для более 500 фильмов
  www.dailyscript.com, www.movie-page.com, www.weeklyscript.com …
- Субтитры (со временем) есть почти для всех фильмов
- Можем сопоставить на основе этой информации



Оценка качества

- Аннотация действий текстом
- Автоматические сопоставление сценария и видео
- Необходимо проверять соответствие



a: - качество сопоставления

Пример ошибки

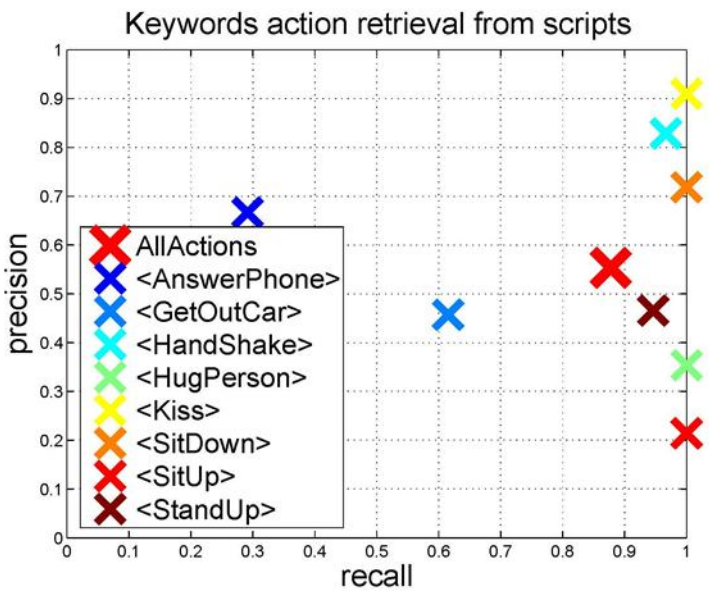
A black car pulls up, two army officers get out.

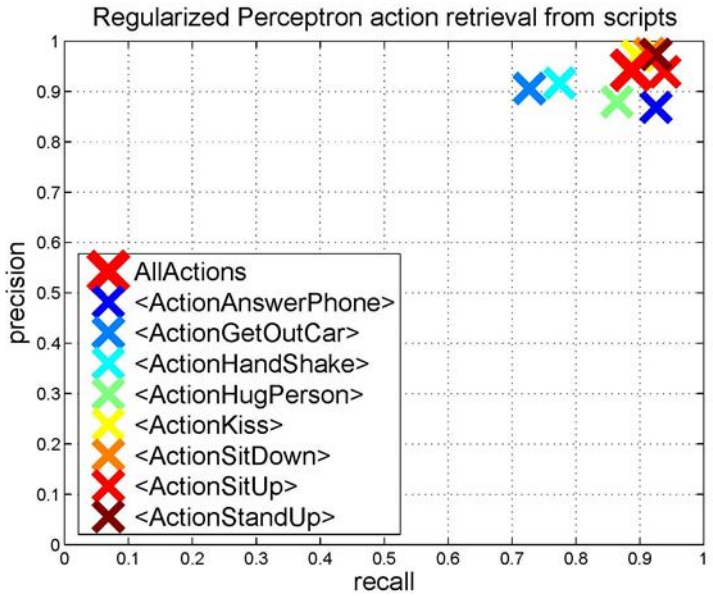
Извлечение действий из сценария

- Высокая вариабельность описаний в тексте:

GetOutCar действий	*"… Will gets out of the Chevrolet. …"* *"… Erin exits her new truck…"*
Потенциальная ошибка:	*"…About to sit down, he freezes…"*

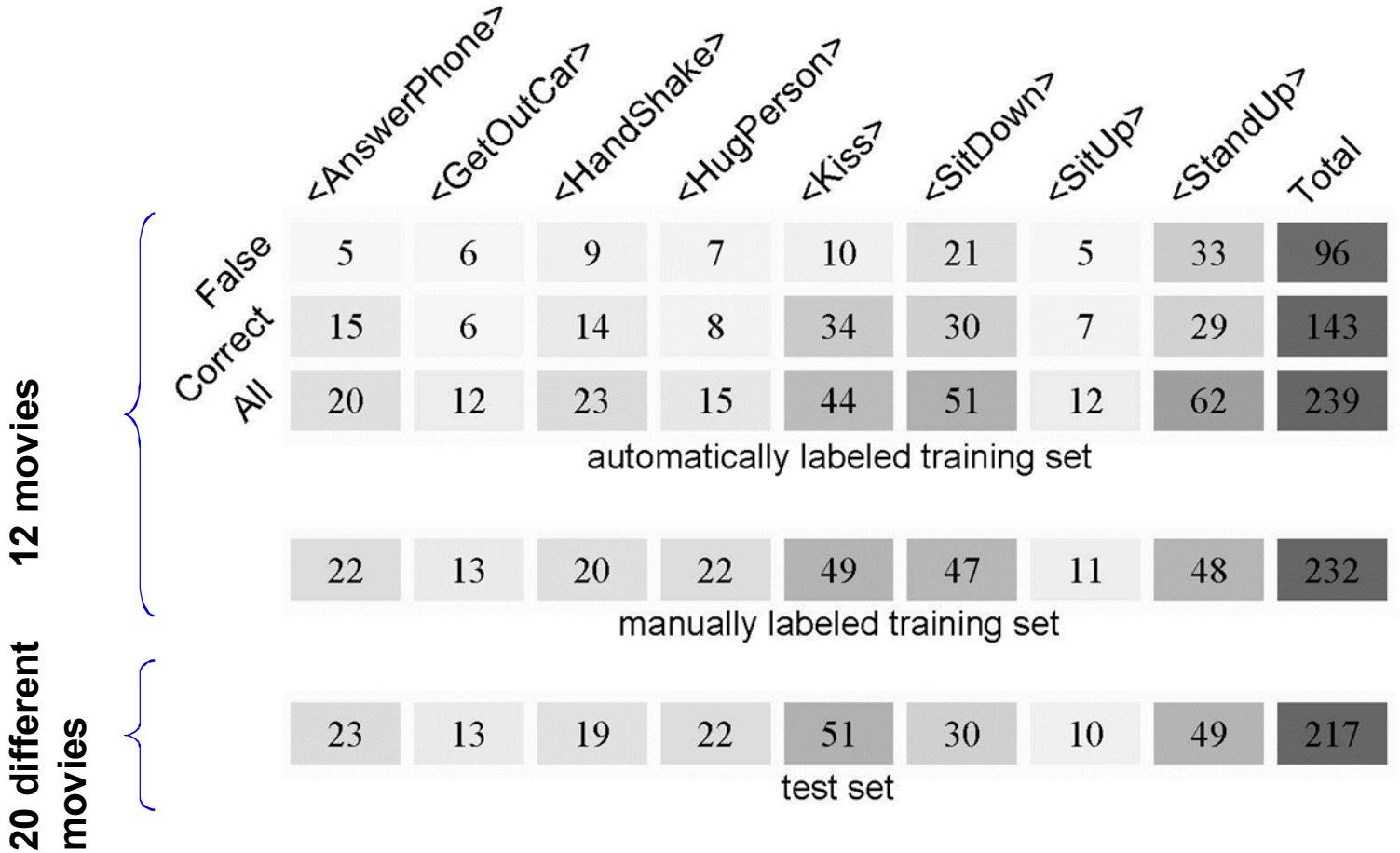
- => Классификация текста с учителем





Набор данных из кино

Microsoft Research



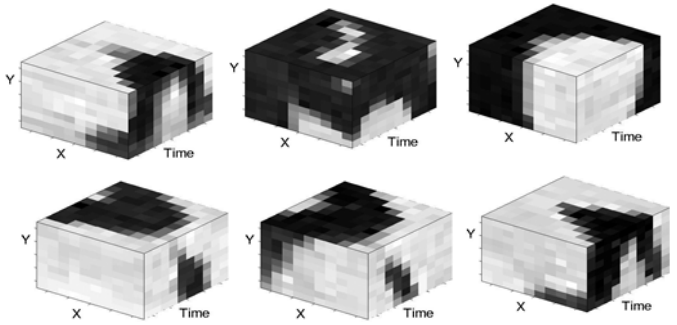
- Обучить классификатор по автоматической разметке
- Сравнить работы с ручной разметкой

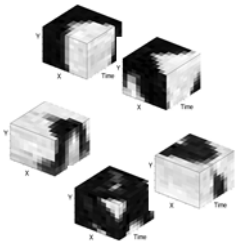
Схем метода

Мешок STIP + многоканальный SVM

[Schuldt'04, Niebles'06, Zhang'07]

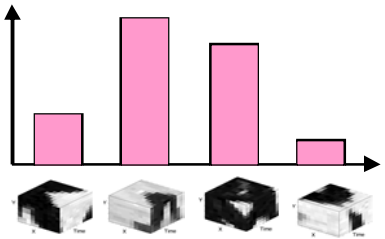
Набор STIP-патчей





HOG & HOF patch descriptors

Histogram of visual words

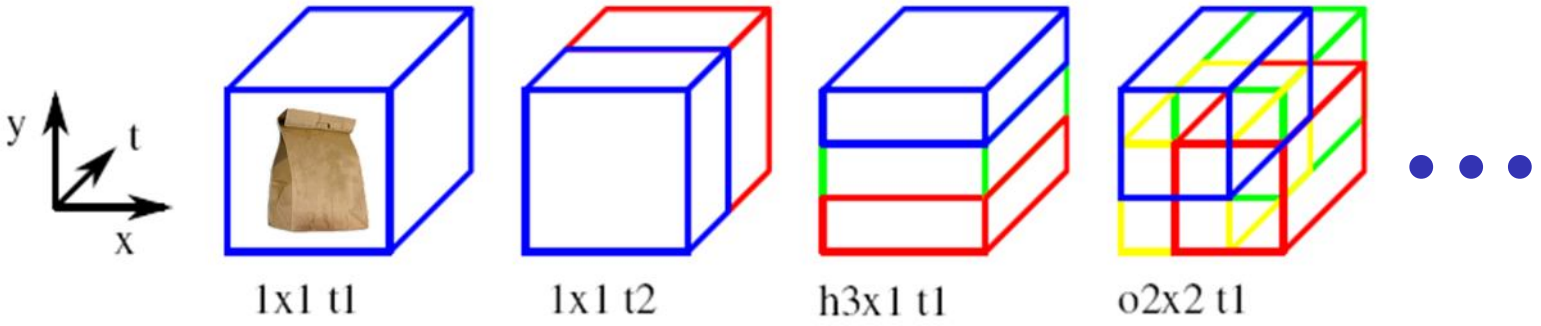


Multi-channel SVM Classifier

«Мешок слов»

Используем разные решетки

- По пространству
  - 1x1 (стандартный мешок)
  - 2x2, o2x2 (50% перекрытие)
  - h3x1 (горизонтальный), v1x3 (вертикальный)
  - 3x3
- По времени:
  - t1 (стандартный мешок), t2, t3



Многоканальное ядро

Microsoft Research

Используют SVM с многокальным хи-квадрат ядром для классификации:

$$K(H_i, H_j) = \exp\Big(-\sum_{c \in \mathcal{C}} \frac{1}{A_c} D_c(H_i, H_j)\Big)$$

- Канал *c* – это комбинация детектора, дескриптора и вида сетки
- $D_c(H_i, H_j)$ – расстояние хи-квадрат между гистограммами
- A_c среднее расстояние между всеми обучающими примерами
- Выбор наилучшей комбинации каналов осуществляется жадным методом

Объединяем каналы

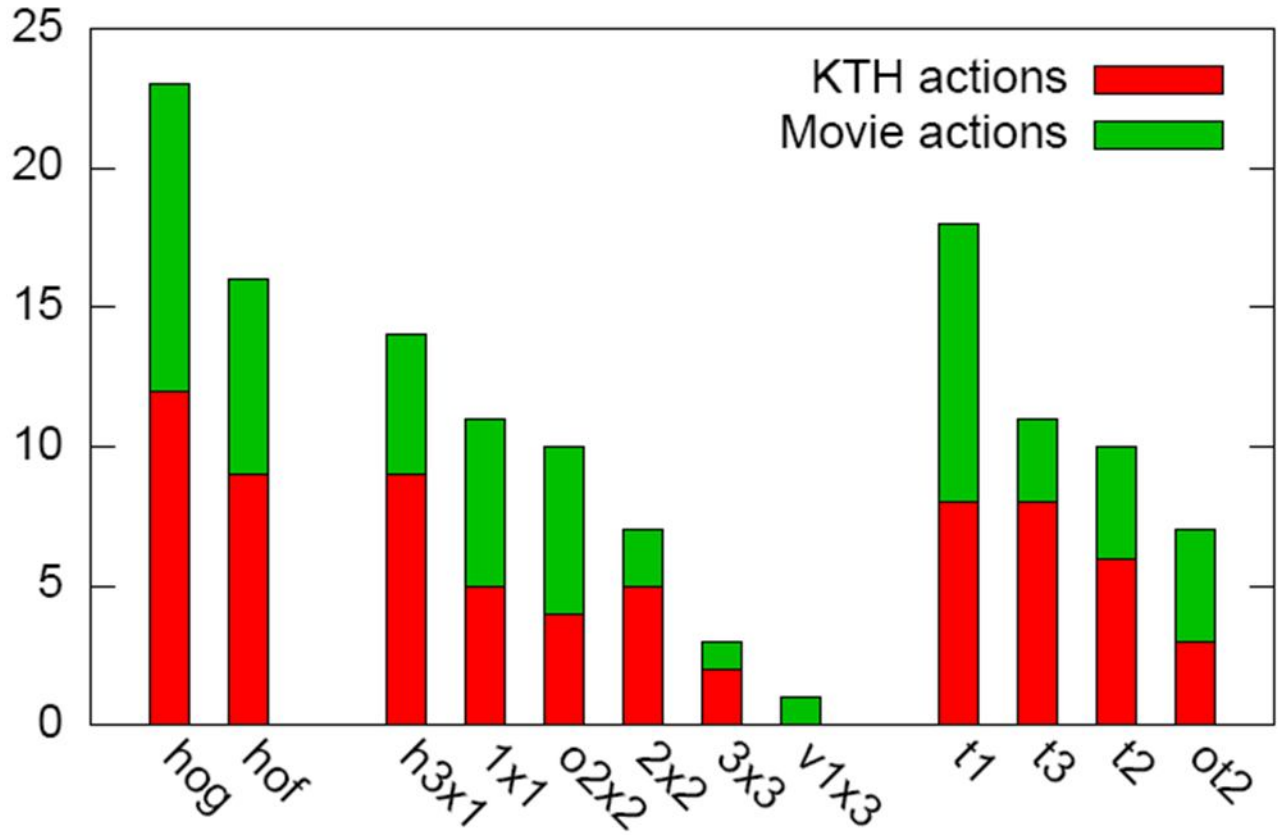
Microsoft Research

Task	HoG BoF	HoF BoF	Best chan.	Best comb.
KTH multi-class	81.6%	89.7%	91.1%	91.8%
Action AnswerPhone	13.4%	24.6%	26.7%	32.1%
Action GetOutCar	21.9%	14.9%	22.5%	41.5%
Action HandShake	18.6%	12.1%	23.7%	32.3%
Action HugPerson	29.1%	17.4%	34.9%	40.6%
Action Kiss	52.0%	36.5%	52.0%	53.3%
Action SitDown	29.1%	20.7%	37.8%	38.6%
Action SitUp	6.5%	5.7%	15.2%	18.2%
Action StandUp	45.4%	40.0%	45.4%	50.5%

Сравнение разных комбинаций

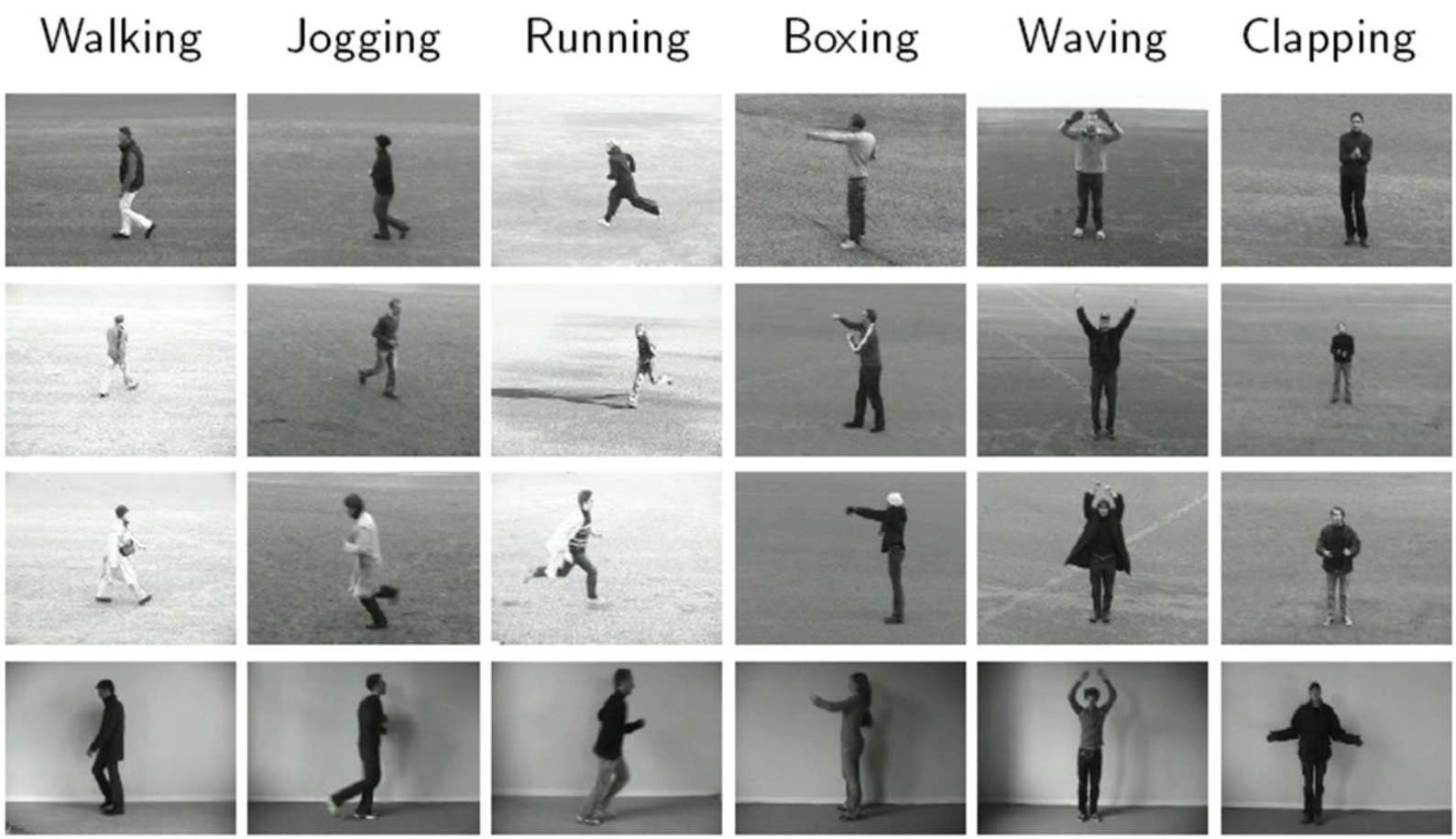
- Разные сетки и комбинации каналов обеспечивают прирост качества

Сравнение сеток



Число использований каждого канала в наилучших комбинациях

Сравнение



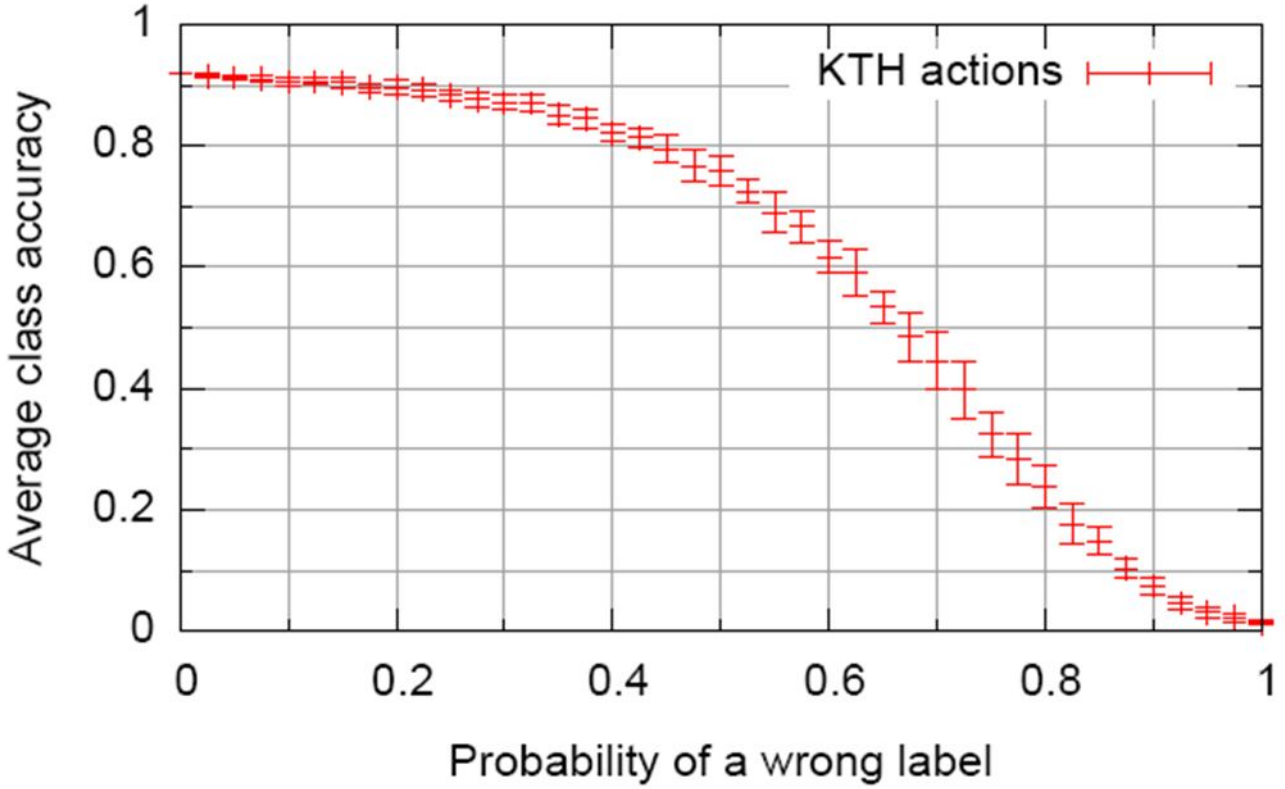
Сравнение

Method	Schuldt et al.	Niebles et al.	Wong et al.	Nowozin et al.	ours
Accuracy	71.7%	81.5%	86.7%	87.0%	

Асредние точности на базе KTH

	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	.99	.01	.00	.00	.00	.00
Jogging	.04	.89	.07	.00	.00	.00
Running	.01	.19	.80	.00	.00	.00
Boxing	.00	.00	.00	.97	.00	.03
Waving	.00	.00	.00	.00	.91	.09
Clapping	.00	.00	.00	.05	.00	.95

Устойчивость к шуму



Деградация качества в присутствии неправильных меток

- До 20% неправильных – незначительное снижение качества
- При 40% - снижение качества на 10%

Распознавание в кино

Примеры работы на данных из кинофильмов

Распознавание в кино

Примеры работы на данных из кинофильмов

Microsoft Research

Распознавание в кино

	Clean	Automatic	Chance
AnswerPhone	32.1%	16.4%	10.6%
GetOutCar	41.5%	16.4%	6.0%
HandShake	32.3%	9.9%	8.8%
HugPerson	40.6%	26.8%	10.1%
Kiss	53.3%	45.1%	23.5%
SitDown	38.6%	24.8%	13.8%
SitUp	18.2%	10.4%	4.6%
StandUp	50.5%	33.6%	22.6%

Сравнение по средней точности для каждого класса при автоматической и чистой (ручной) разметке данных

Microsoft Research

Пример

Microsoft Research

Резюме лекции

- Оптический поток – основной источник информации о движении в сцене, один из базовых инструментов для компьютерного зрения
- Для распознавания видео мы можем использовать те же подходы, что и к изображению, но переведя их в трёхмерные пространственно-временной объём
  - Скользящее окно
  - Особенности, детекторы и дескрипторы
  - Мешок слов и методы классификации
- Данные и их объём является общей проблемой. Нужно использовать «слабую» разметку.