E6876 Project Proposal
Object tracking with robust PCA initialization and sparse recovery

Yanchen Liu yl4189, Chenye Yang cy2540

1. Introduction

1.1. Background and related works

The Olympic Games is a well-known event, and every time being held, it will attract the attention of hundreds of millions of people worldwide. Apart from that, there are lots of high-profile sports events, such as the FIFA World Cup, FIA Formula 1 World Championship and Super Bowl. Most sports require athletes to move at a very high speed, such as football, racing, diving. To ensure the audience a good watching experience, the athletes must be placed in the center of screen and the video stream should have a high quality.

To achieve those two goals above, the problems to solve are as follows:

1) Track the athletes in a video stream robustly in real time
2) Compress the raw video stream and recover a high-quality video

There exist many methods for object detection, such as the R-CNN family and YOLO family. To combine with the knowledge in class, we decide to use:

1) Robust PCA, to do background segmentation as initialization
2) L1 tracking, to do object detection and tracking
3) Compressed sensing, to do image sampling and reconstruction

1.2. Video data description

These two videos are used as the test for our method:

1) Motorcycle Stunt Show first class from Motor Bike Expo 2016. In this video our aim is to separate and track the motor and its player in the moving

situation and try to compress and reconstruct a high resolution video based on a low quality video.

2) Daily walking video from us. In this video, one person walks in a normal daily environment with some obstructions blocking between this person and camera. This video has an ultra-high 4K resolution and is obtained by a normal smartphone.

2. Workflow and algorithms

2.1. Workflow

For the motor video, we would like to apply robust PCA to get the foreground as the initialization input for the L1 tracker. Then as an independent part, compress the video and recover it.

2.2. Algorithms

2.2.1. Robust PCA

Like classic PCA, Robust PCA (Robust Principal Component Analysis) is essentially the problem of finding the best projection of data in a low-dimensional space. Robust PCA considers such a problem: the general data matrix D contains structural information and also contains noise. Then you can decompose this matrix into two matrices and add them: $D = A + E$, A is low rank (due to certain internal structural information causing linear correlation between rows or columns), E is sparse (containing noise , which is sparse), then Robust PCA can be written as the following optimization problem:

$$min_{A,E}\ rank(A) + \lambda ||E||_0\ \ s.t.\ A + E = D$$

Because rank and L0 norms have non-convex and non-smooth characteristics in optimization, this NP problem is generally converted into solving a relaxed convex optimization problem

$$min_{A,E}\ ||A||_* + \lambda ||E||_1\ \ s.t.\ A + E = D$$

2.2.2. L1-minimization tracking

The main idea of L1 Tracker is to use the images (features) obtained in the first frame and the last few frames as a dictionary, and then add a lot of trivial templates, that is, there is only one white pixel in an image, and the rest are black. In this way, the newly obtained particle filtered candidate image is projected onto this set of dictionaries using L1 least squares criterion. If the coefficients of the first few templates are relatively large, then it may be the target to be tracked. If there are many trivial templates with relatively large coefficients, then it may be the background.

2.2.3. Single-frame image super-resolution reconstruction

Suppose x is the image with high-resolution, y is the image with low-resolution. Also for any image, $x \ or \ y = D \times \alpha$. Then we have that:

$$y = A \times x = A \times D \times \alpha$$

where A is the sensing matrix, D is the over-complete dictionary for high-dimensional data, α is the coefficient vector. Our goal is to train the D and calculate α.

Our training data is high-resolution video stream. After degeneration, each frame of the video,which is a high-resolution image, forms a low-resolution image. These images are used to train the dictionary matrix D_h, D_l. D_h, D_l are dictionaries for high-resolution and low-resolution images correspondingly:

$$x = D_h \times \alpha_h \, , \; y = D_l \times \alpha_l$$

According to the compressed sensing theory, once we know the coefficient vector α of low-resolution image to dictionary D_l, we are able to recover the corresponding high-resolution image. The coefficient vector is calculated as follow:

$$\min_{\alpha} \gamma\|\alpha\|_1 + \tfrac{1}{2}\|FD_l\alpha - Fy\|_2{}^2$$

$$\min_{\alpha} \gamma\|\alpha\|_1 + \tfrac{1}{2}\|FD_l\alpha - Fy\|_2{}^2 + \tfrac{1}{2}\|PD_h\alpha - \omega\|_2{}^2$$

where matrix P extract the overlay area of two adjacent tiles, ω is the value of the overlay area, F is the high pass filter operator.

According to the dictionary of high-resolution image D_h and the coefficient vector of low-resolution image α_l, the high-resolution image can be recovered:

$$x = D_h \times \alpha$$

However, there may be over-recovery or under-recovery. Thus we need to add a global constraint to the recovered image x_0 to get a better x:

$$x = \arg\min_x \|x - x_0\|,\ s.t.\ y = A \times x$$

2.3. Evaluation

For the robust PCA, we use intersection over union (IoU) as a quality mark. Since the particle filter's output is highly related to the selection of the initial box, if IoU of robust PCA is lower than 50%, we will apply human label for the next stage of the workflow.

For the L1-minimization tracking, since it involves all frames in the video and we want to ensure the processing time is lower than a threshold, we will sample 10% of the whole video frames randomly and calculate the mean IoU as the evaluation index for the tracker.

For video compressing and super resolution, we use the MSE (Mean Squared Error) or compression ratio as the evaluation.

3. Division of work

Yanchen Liu: Robust PCA, L1-minimization tracking
Chenye Yang: Single-frame image super-resolution reconstruction

4. Reference

X. Mei and H. Ling. Robust visual tracking using l1 minimization. In ICCV, 2009.

J. Yang, J. Wright, T. S. Huang and Y. Ma, "Image Super-Resolution Via Sparse Representation," in IEEE Transactions on Image Processing, vol. 19, no. 11, pp. 2861-2873, Nov. 2010.