International Computer Vision Summer School

ESSAY COMPETITION - The Social Impact of Computer Vision

Essay Topic:

Assistive computer vision

(ten years of technological transfer and research with future prospects)

Author: Behzad Mirmahboub

University of Genova

Italian Institute of Technology (iit)

behzad.mirmahboub@iit.it

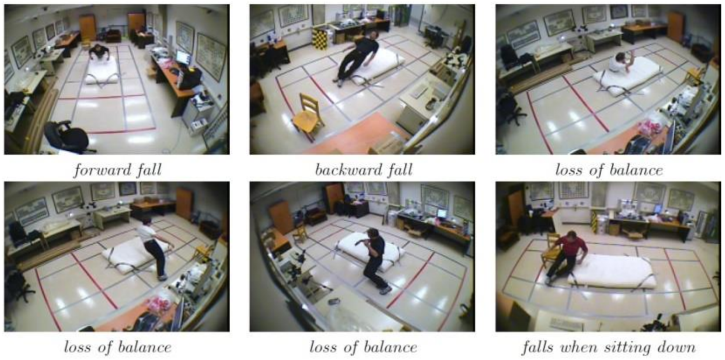
June 2016

Introduction

“According to one well-known story, in 1966, Marvin Minsky at MIT asked his undergraduate student Gerald Jay Sussman to spend the summer linking a camera to a computer and getting the computer to describe what it saw. We now know that the problem is slightly more difficult than that.” [Computer Vision, Algorithms and Applications, Richard Szeliski, 2011]. Today advances in computer algorithms such as deep networks and increased power of computations using GPUs have caused some of our dreams in computer vision get close to reality. In this essay I will briefly explain about two topics of “human fall detection” and “human re-identification” that I worked on before.

Human Fall Detection

Nowadays improvements of life conditions and better health care systems have increased the average human life. Population of old people is more than before. Many of these persons live alone far from their children. Unusual events such as falling on the ground may happen for these seniors that will lead to serious injuries with painful consequences. Therefore they need continuous health care at their own home. Some persons hire full-time or part-time nurses to help them at home. But it is expensive and also restricts their privacy and is not possible for all.


forward fall *backward fall* *loss of balance*

loss of balance *loss of balance* *falls when sitting down*

Figure 1: Examples of falling [2]

One solution is to use assistive wearable equipment that continuously monitor the person conditions and produce alarm in case of abnormality. This sort of equipment consists of accelerometers or gyroscopes to find the person situation. But using these wearable devices is boring in long time and many persons don’t like to use them continuously.

Another solution is to install surveillance cameras inside house to monitor the persons using some qualified staff outside the house. This creates unsatisfying condition that puts the privacy of habitants in danger. Recently a lot of efforts have been done in development of a computer vision algorithm to automatically detect abnormal conditions such as serious falling on the ground. The advantage of this algorithm is that it processes the video directly and don’t store or send the original videos. Therefore the privacy of the user is not in danger. Such an algorithm can be implemented on hardware inside a specialized camera. The output of that camera will be just some features that are extracted from video and no original video will be produced. The extracted features are some numbers that are used for event classification and don’t reveal any information about the identity of the related persons. If an abnormal event is detected base on the extracted feature, then an alarm will be produced and will inform the authorities for immediate emergency actions. An ideal vison based fall detection algorithm should discriminate between abnormal events and usual day life activities such as sitting and sleeping.

Human Re-Identification

Security in crowded environments such as train stations, airports and markets is increasingly important now. Many policemen and soldiers are hired to control populated areas. Surveillance cameras are installed everywhere to monitor the people. Such a camera network needs an intelligent center to watch and interpret the videos. This is an extremely time consuming task and is also prone to human errors.

Here computer vision algorithm can be helpful. The idea is to have a dataset of people's image that are captured in one camera. The goal is to recognize a person who is appeared in another camera as shown in Figure 2. Matching individuals between different non-overlapped cameras is still an open issue because of several hard challenges.

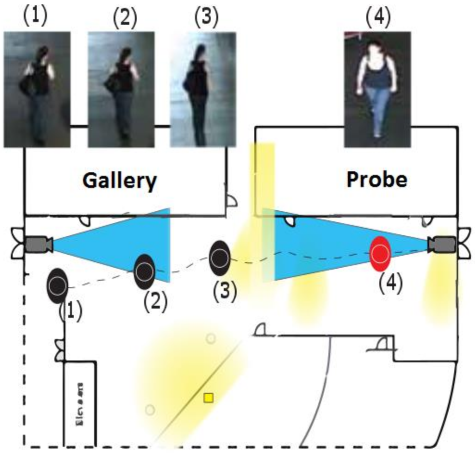


Figure 2: Camera layout for re-identification [4]

Firstly, the images are captured in uncontrolled environment with large field of view and low resolution. Therefore some biometric features such as person face that depend on image details are not accessible. Visual appearance features such as clothing color and texture are the main features that are used in human re-identification. The main assumption is that the persons will not change their clothing while they pass from one camera to another one. Images that are captured with different cameras are affected by variation in illumination, human pose, different backgrounds and occlusions. Therefore the discriminative features such as color histogram that are extracted from them are drastically different. One possible simplification is to consider the required time for a person who travels from one camera to another one based on their distances and to remove some persons that are not possible to appear in specific cameras.

Person re-identification methods are roughly divided into two groups of feature design and metric learning. The methods that concentrate on feature design try to extract discriminative descriptors from images that robustly identify persons. On the other hand, metric learning methods take a descriptor vector and try to find a weighting matrix that emphasizes the important elements of the descriptor when they want to calculate the distance between two descriptors.

Re-identification is a relatively new problem, far from being solved definitively, and strongly dependent on the quality of the used sensory data. The solution of this problem can be divided into three levels of difficulties: the short term (about 2-3 years), medium term (about 5 years) and longer term. Research challenges are very dependent upon the quality of the input data.

1- Short-term Challenges

In this case, high quality input images are available without significant occlusions and background clutters. The scenarios are rather constrained and cooperative. Constrained means that the people are limited to certain regions (e.g., walkways) and their features can be identified and extracted. Cooperative means that they are not trying to evade identification. This means that they are willing to provide their images in a fixed setup (e.g., a security checkpoint).

In such conditions, the pre-processing part of the re-id process is relatively reliable, that is, detection and tracking of moving objects is possible. Foreground can be separated effectively from background and body extraction can be done. Therefore features can be extracted to build a meaningful descriptor to do consequent actual recognition.

2- Medium-term Challenges

Here input images have slightly lower quality, with some partial occlusions. Persons still behave in a cooperative but relatively unconstrained way. The problem here needs a longer-term investigation. Examples can be people walking on a not-too-busy street and not objecting to having their images taken.

When there are partial occlusions and appearance variations in input images, extracting discriminative features is very challenging. In such cases, detection and tracing are not reliable and more roust descriptors such as 3D features are necessary. Considering attributes of the image may be useful. Attributes are semantic mid-level features in the image (e.g. a person who carry a bag with him/her-self).

Many re-identification methods are designed to recognize person's images between two cameras. But these methods may be inconsistent when we deal with three or more cameras. Another simplification assumption in most current re-identification methods is that they consider closed-world scenarios. It means that each person appears in both cameras. In the other words for each person in each camera there is a match in another camera. This assumption is not true in real world scenarios that are called open-world. Persons may appear or disappear in one camera without a match in another one. Currently most of the methods works with images datasets from fixed cameras that may not be hold in real world.

3- Long-term Challenges

This is the real-world scenarios consist of low-quality images, unconstrained and uncooperative conditions with challenges that will need a longer time horizon. This will involve dealing with natural videos with high clutter in the data, and severe variations in the environmental conditions. An example could be a busy city scene, where a person needs to be recognized as he walks through several blocks with large blind areas in between.

More robust feature extraction is needed in this case. Some tools such as image restoration can improve image quality. Semantic attributes can helps to design more robust features. Some techniques such as deep learning and sparse coding can automatically learn the best features instead of hand-engineered ones. Data from other sensors such as infra-red camera rather than optical camera can be used in order to find salient part of the persons in the image.

Traditional methods of person re-identification that use supervised learning, try to train a model with as much labeled images as possible to account for human pose variations. In addition to expensive human labor that is needed for annotation, such static models that are trained on limited datasets are not scalable to other datasets. Some methods try to reduce human effort for labeling the samples. Semi-supervised or unsupervised learning have been proposed by some works that try to exploit saliency or weight features to improve re-identification. Active learning tries to involve human in labeling only for difficult samples. Domain adaptation techniques train a model on one dataset and try to use it on other related datasets.

The assumption of unchanged clothing of human may not be hold in long-term person re-identification. Some 3D soft-biometric cues such as body part length and their ratios can be considered.

References

[1] Behzad Mirmahboub, Shadrokh Samavi, Nader Karimi, and Shahram Shirani. "Automatic monocular system for human fall detection based on variations in silhouette area." IEEE Transactions on Biomedical Engineering, vol. 60, no. 2, 2013, 427-436.

[2] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. "Multiple cameras fall dataset." DIRO-Université de Montréal, Tech. Rep 1350 (2010).

[3] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. Person re-identification. Vol. 1. London: Springer, 2014.

[4] Amran Bhuiyan, Alessandro Perina, and Vittorio Murino. "Exploiting multiple detections to learn robust brightness transfer functions in re-identification systems." IEEE International Conference on Image Processing (ICIP), 2015, pp. 2329-2333.