View-invariant video analysis

Patrick Pérez,
Technicolor Research and Innovation, France

Abstract

Effective annotation, browsing and search of video contents, either professional or user generated, will require in the future a number of automatic and semi-automatic analysis tools. Apart from generic image analysis and scene understanding tasks (image segmentation, object detection and categorization, etc.), some tasks are specific to video in that they rely on spatiotemporal information. These tasks include shot detection, detection of moving objects, motion-based segmentation, visual tracking, motion characterization, human action recognition, etc.

As most other image analysis tasks, these video analysis tasks require to compare, match, cluster and recognize pieces of visual information in presence of various types of appearance changes. When tracking the position of an object in the camera field of view, for instance, the size, shape, color and texture of the object might change dramatically through time due to its relative motion w.r.t. the camera and the light sources, and to its own dynamic deformations.

When building models and tools to address video analysis tasks, the dependency with respect to time-varying view and scene parameters must thus be addressed one way or another. In particular, a certain degree of invariance with respect to some of these parameters is desirable. There are different ways to address this issue:

1. Explicit modeling of appearance variations as a function of several parameters related to the configuration of the objects, the scene and the camera set-up.
2. Sampling of the appearance space irrespective of explaining parameters.
3. Design of matching metrics that ensure a certain amount of view invariance despite such invariance not being built in the descriptions they rely on.
4. Design of appearance descriptions that are invariant, to a certain degree, to view point and scene configuration.

Depending on context, one of these approaches might be more adapted or simply more practical than the others. When aiming at image-based motion capture for instance, the 3D articulated pose of the body is not a nuisance parameter but the main information to be estimated. In that case, approaches of type (1) are preferred. On the other hand, if action recognition is the aim of the analysis, such detailed pose parameters, whose estimation from monocular videos is difficult, could be ignored with approaches of type (2), if multi-view data collection is possible, of types (3) or (4) otherwise.

Approaches of type (1) will appear in different guises in some of the lectures of the Summer School. In this lecture, we shall focus on approaches of types (2), (3) and (4), by which visual descriptions and comparisons are proposed for different video analysis tasks, with no modeling of the dependency on pose and viewpoints. This invariance comes at the cost of reduced discrimination power, and a compromise has to be found. We shall investigate such approaches in the context of three different tasks: (1) robust visual tracking; (2) temporal synchronization of different views of a same event or of similar actions; (3) recognition of human actions. To this end, various low-level image and video descriptors will be mobilized, including color and gradient histograms, point tracks, optical flows and temporal self-similarity.

Syllabus: video analysis; view invariance; low-level spatial and spatiotemporal descriptors; temporal self-similarity; visual tracking; view-invariant temporal video alignment; action synchronization and recognition.