Capabilities and Limitations of Visual Surveillance

Ingo Lütkebohle
iluetkeb@techfak.uni-bielefeld.de

Faculty of Technology, Applied Computer Science
Bielefeld University, P.O. Box 100131, 33501 Bielefeld, Germany

1 Introduction

Surveillance cameras have become widespread: Public places, shopping centers, offices, transportation, and the list could go on. Increasingly, the sheer volume of data requires automated analysis. This poses serious questions: Surveillance is touted as a tool against crime but *does it really work?* On the other hand, it is feared that we might wake up in a world where our every step is being monitored and scrutinized. *How close are we to 1984s Big Brother?*

Of course, no definite answers will be forthcoming. However, a review of the technology that drives automated surveillance systems may shed some light on what it can and cannot do. The technology has been making great strides in recent years but quite a few problems have only been sidestepped. Sometimes, these problems are very revealing, hinting at fundamentally hard problems.

To start this off, some general remarks on the workings of visual observation are provided in the remainder of this section. Part 2 will review methods of automated visual surveillance, from feature extraction to recognition and discuss capabilities and limitations throughout. The conclusion in 3 delivers a high-level view of capabilities and limitations.

Marvin Minsky once stated that "In general, we're least aware of what our minds do best" [13], referring to the fact that many of the things humans consider 'easy' just appear that way because we learned them so well. Their full complexity becomes evident, however, when trying to build automated systems. Visual observation is such a task: We can effortlessly tell what other persons are doing just from looking at it, right? Well, not quite, but even where that is true, automated systems are still far from being able to do the same and its not from lack of trying by the designers!

Furthermore, Minskys statement has a second part to it: We often don't know *how* we accomplish the "simple" things. For instance, and contrary to common belief, our powers of visual observation may not be learned from visual experience alone. Recently, a number of psychological findings suggest that the *motor experience* we have from our own body is at least as, if not more, important (e.g., compare [10, 2]). Therefore, **purely visual analysis may not be enough** and external knowledge will still be necessary. As such knowledge comes from human designers, it is a crucial limiting factor to the capabilities of an automated system.

2 Intro to Visual Analysis

2.1 Overview

In the most basic view, visual analysis starts with a camera and ends with action recognition. On closer look, it rapidly starts to become complicated and thick books have been written on just small parts of the problem. One of the most current, accessible and self-contained textbooks is "Computer Vision: A Modern Approach" by D.A. Forsyth and J. Ponce [8] and I will point out pertinent chapters where appropriate throughout this section.

Visual Surveillance Systems Though details vary a lot, most vision systems contain a processing pipeline at their core, as shown in figure 1. The separation is mostly due to different algorithms.

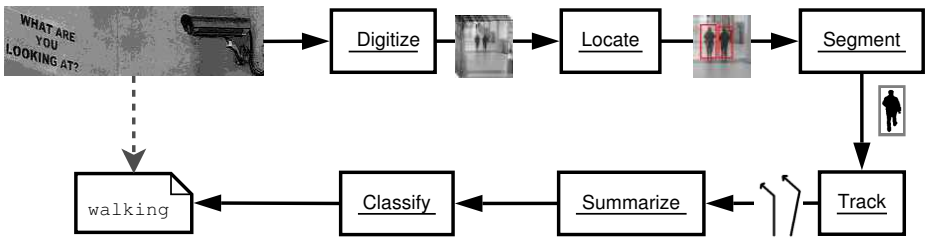


Fig. 1. Example of surveillance analysis

For a "real-world" system, essential additions to this sketch would include at least some form of storage and retrieval system, multiple camera inputs and algorithm redundancy. Common are active camera control (for pan, zoom and tilt) and feedback mechanisms that can tailor the performance of early-stage algorithms to the current analysis task.

The system shown could be an example of the stages for person tracking. The general ideas are simple: First, the object of interest must be located and described. Here, these are two persons and the description is their shape image. This is performed over several frames with data-association ("tracking"), yielding related sets. In this case, two trajectories in time. These are then summarized, depending on the needs of the final algorithm, and classified (by comparison to prior examples), e.g. as "walking" (as opposed to window-gazing or something the like).

For each step, many methods and algorithms exist. The remainder of this section will give a short overview and provide some references for further reading.

[0] Camshot by Bhikku (postprocessed): http://flickr.com/photos/bhikku/1187679/
Surveillance images from CAVIAR: http://groups.inf.ed.ac.uk/vision/CAVIAR/

2.2 Locating Humans

As shown in figure 1, the first step is locating the interesting parts in the frame. Surveillance systems are looking for humans, a hard problem because of considerable variability in appearance (shape, color and texture).

Temporal Differencing A trivial approach to detect humans is to look for any changes in the image from one frame to the next. When the camera is not moving, temporal differencing is an incredibly easy approach to do so: Take two frames, subtract pixel-by-pixel, apply a threshold, see image to the right...

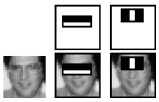
Because of the frame-by-frame approach, it is very adaptive to changes in the environment and another advantage is that it does not make assumptions about the scene. However, it can be problematic that only motion at edges is visible for homogeneous objects (this particular problem has been tackled under the heading of "background subtraction", for instance cf. [14, 6]). For outdoor use, even a small camera shake – e.g., because of wind – can also cause failure if not compensated for. Therefore, this is usually just one component of a larger system.

Optical Flow Considerably more powerful is to assign a motion vector to every pixel of the image by comparison of successive frames. First popularized by Horn & Schunk [9] for detecting ego-motion of robots, in which case the camera is moving (hence the name – when the robot and its camera moves, the rest of the image "flows by"). It has a number of useful properties for this case and makes good use of low-resolution images by smoothing over all pixels for often impressive accuracy.

However, with the static camera setup typical for surveillance, troublesome edge cases are more frequent, such as overlapping objects moving in different directions and also the general "aperture problem" that motion is only unambiguous at corners. A good solution is therefore difficult to find or, in effect, optical flow is either *slow* or *inaccurate*. For instance, an accurate state-of-the art optical flow algorithm by Bruhn et al [3], achieves 18fps on a 316x252 sequence using a 3GHz Pentium 4, which is considerably slower than most of the other approaches presented. The result is very detailed, but comes at a high price.

Skin Color While this may sound particularly silly, given the huge natural variation, "skin color" detection is a fairly common approach, mostly used in conjunction with other cues. For instance, given a body silhouette, skin color may be a good cue to find hands and face. A fairly recent and comprehensive comparison of skin color matching [11] uses images on the web to gather a large data-set (with some fairly obvious results of questionable generality). Of course, this approach is only applicable for surveillance systems that use color cameras.

Appearance As human appearance is very diverse, its direct use for detection was restricted to very limited applications for a long time. In the last years an approach based on an *automatically selected combination of very simple features* ("boosting") has made great progress and is now the reigning champion for body part detection, especially face detection. To give an idea of just how simple the features can be, the original example due to Viola & Jones [17] is shown on the right.



Much more than these two are needed to be robust, however – typically, between one and two thousand features are combined, in a coarse to fine cascade that applies subsequent features only when the first matches.

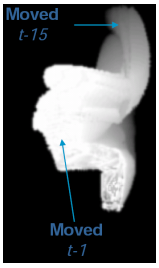
Appearance based detection requires sufficient resolution to disambiguate body-parts. This is usually possible from surveillance data but crowd shots (such as at sports events) do not suffice. Apart from that, a problem of this approach is the amount of training data required, see 2.4. Its major advantages are that it can detect humans without motion, provides a very precise localization of individual parts and makes very few mistakes of the sort that something which is not a human is mistakenly detected as one (false positives). Very recently it has also been extended to articulated body parts such as hands (e.g. [12]).

2.3 Intermediate descriptions (Summaries)

The methods reviewed so far indicate a number of possible locations for humans in an image. They are not yet suitable for further processing however, because: a) multiple humans might be present, possibly overlapping and b) they are still basically at the level of a single frame. The task of separating them is called *segmentation* and will only be touched upon here, then various methods for summarization in time or space will be described in more detail.

Segmentation One of the most ill-defined topics of computer vision, segmentation is a problematic area mostly because everyone seems to expect different things from it. An example for surveillance is whether to segment the human as a whole or whether to identify individual body parts, too. That said, segmentation is usually performed based on a combination of the cues reviewed: For example, spatially connected regions of similar color moving in the same direction are good candidates for segmentation from their neighbors. For more information, please see [8, chapters 14-16] and also the discussion in part 3.

Motion History Based on image differencing, motion history images [5] accumulate differences over multiple frames into a single overlayed image. They are inspired by human peripheral vision and especially suited for capturing large body motions. The image on the right[1] shows a side view of a person sitting. Brighter parts of the image represent newer information, so that sitting down and standing up can be distinguished.



[1] A. Bobick, *Cognitive Vision Summer School 2005.*

Trajectories Trajectories abstract from appearance and capture position over time only. They can be represented visually (see the white trail in the image on the right[2]) or as paths with a given direction, velocity and duration. Input for trajectories can come from skin color detection, appearance detection and so on. It should be noted that tracking becomes a non-trivial problem quickly, such as with many different motions at the same time, occlusions (due to other people or objects) and the like.

For cases such as single walking, linear tracking using Kalman filters [8, chapter 17] works well. When overlap or interaction occur, however, some form of disambiguation by appearance will be required. Furthermore, any actions that severely change body form (such as bending down or dancing) are out of scope for regular tracking and require specialized methods such as particle filters [7], which keep track of the different body parts separately but are not yet ready for production use.

Posture A mid-level description between trajectories and full limb-tracking is the posture. On its own it is capable of distinguishing between different ways to execute the same action, e.g. compare [1] for distinguishing between different ways of walking (Canadian laws on liability for drunken behavior make this interesting). Also, gait analysis based on posture makes a medium-range biometric [4].

One issue with silhouette-based posture is self-occlusion (compare right arm in the silhouette image[3]). It is also often context-dependent and needs to be complemented with a sequence-based recognition method.

2.4 Recognition

Recognition is the process of assigning a human-understandable label to the data and is usually performed with methods from **machine learning**. The challenge is that explicit specifications of *how* to perform recognition are next to impossible and work that tries to define classes automatically has not produced anything close to human intuition so far. Therefore, all of the following methods work from examples, which is known as **supervised learning** and works as follows: First, humans have to manually assign the desired labels to input sequences, creating classes. The combination of data and labels constitute the **training set**. It is presented to the recognition algorithm, which tries to find structure that is distinctive for one of the given labels. During production, new data comes in from pre-processing and is assigned to the best match amongst the learned classes. Voilà, Recognition! Some, but not all!, algorithms can also reject input that does not fit any of the learned classes.

[2] A. Bobick, *Cognitive Vision Summer School 2005*
[3] Courtesy of C. Bauckhage [1]

A thorough review of learning methods is unfortunately beyond the scope of this paper (its long enough as it is). The literature mentioned so far, and especially the book by Forsyth & Ponce [8, part IV and VI] includes material on relevant recognition methods. All I will try here is to convey a rough idea of which method to choose for what kind of problem.

For the simplest method is euclidean distance or normalized correlation [8, section 7.6]. Principle Component Analysis (PCA) generalizes both to more examples. Both of these can be used with Bayesian classification, which has been done for many things from background subtraction to face recognition (cf. [16]). Better performance is often achieved by "boosting" or Support Vector Machines (SVMs). For an overview of the various methods, see [8, chapter 22].

Whenever the sequence to be classified consists of multiple examples over time where the precise duration is not known, Hidden Markov Models (HMMs) or more generally, graphical models are applicable to the problem.

However, for all their power and sometimes astonishing performance, these methods make decisions based on their input only. In other words, they will pass through the problems of the methods reviewed above and differ mainly in how susceptible they are to bad input. This is often summed up as either **it's the feature, stupid** or **garbage in, garbage out**, depending on your viewpoint.

3 Conclusion

Having reviewed a number of techniques for visual surveillance, back to the questions of a) do they perform as designed, b) is that enough and c) should we be worried? Most systems have been designed for controlled conditions only. This reflects the *focus on intrusion detection and collection of evidence for forensic use*. For these situations, while every individual approach has its shortcomings, a combination can ensure good performance.

For all other situations, the basic problem is that appearance is ambiguous. Therefore, no system can learn on its own, it always has to be assisted by humans. As training material costs a lot of money and time to produce, this severely restricts robustness of systems.

Another, more severe, problem with the production of training material is that it can rapidly become out of date. For instance, once the targets of surveillance become aware of how they are picked out, they are likely to change their habits rapidly. To counter this, learning has to become continuous and on-line, which opens up avenues for manipulation by providing bad examples deliberately. Imagine the Surveillance Camera Players [15] done with a slightly different purpose! It is a completely open research question whether this problem is solvable.

Apart from this question, future reviews should concentrate on detailed analysis of activities and especially interaction, two relatively immature but very active fields of research. These fields are more diverse in that they often concentrate on human-computer-interaction, but some of their methods are also applicable for surveillance.

References

1. C. Bauckhage, J. Tsotsos, and F. Bunn. Detecting abnormal gait. In *Proc. Canadian Conf. on Computer and Robot Vision*, pages 282–288. IEEE, 2005.
2. R. Blake, L. Turner, and M. Smoski. Visual recognition of biological motion is impaired in children with autism. *Psychological Science*, 14(2):151–157, 2003.
3. A. Bruhn, J. Weickert, C. Feddern, T. Kohlberger, and C. Schnorr. Variational optical flow computation in real time. *IEEE Transactions on Image Processing*, 14(5):608–615, 2005.
4. R. Collins, R.Gross, and J.Shi. Silhouette-based human identification from body shape and gait. In *Proc. Int. Conf on Automatic Face and Gesture Recognition*, pages 351–356, 2002.
5. J. W. Davis and A. F. Bobick. The representation and recognition of human movement using temporal templates. In *CVPR '97: Proc. Conf. on Computer Vision and Pattern Recognition (CVPR '97)*, page 928. IEEE, 1997.
6. A. Elgammal, D. Hardwood, and L. Davis. Non-parametric model for background subtraction. In *Proc. of the 6th European Conference on Computer Vision (ECCV)*, volume 2, pages 751–767, 2000.
7. D. Forsyth and J. Ponce. Tracking with non-linear dynamic models, 2003. Orphan Chapter from 'Computer Vision, A Modern Approach', http://www.cs.berkeley.edu/ daf/bookpages/pdf/particles.pdf.
8. D. A. Forsyth and J. Ponce. *Computer Vision, A Modern Approach*. Prentice Hall, Upper Saddle River, NJ, USA, 2003.
9. B. Horn. *Robot Vision*. MIT Press, 1986.
10. A. Jacobs, J. Pinto, and M. Shiffrar. Experience, context and the visual perception of human movement. *Journal of Experimental Psychology: Human Perception & Performance*, 30(5):822–835, 2004.
11. M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.
12. M. Kölsch and M. Turk. Fast 2d hand tracking with flocks of features and multi-cue integration. In *Proc. of the IEEE Workshop on Real-Time Vision for Human-Computer Interaction (CVPRW04))*, volume 10, page 158. IEEE, 2004.
13. M. Minsky. *The Society of Mind*. Simon & Schuster, 1988.
14. N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):831–843, 2000.
15. Surveillance Camera Players. Completely distrustful of all government, 12th December 2005. http://www.notbored.org/the-scp.html.
16. M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. IEEE CS Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, 1991.
17. P. Viola and M. Jones. Rapid object detection using a boosted cascase of simple features. In *Proc. IEEE CS Conf. in Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518, 2001.