BMVA Detection vs. Tracking Meeting

Research Summary - Optimal Importance Sampling for Human Motion Tracking using Persistent Low-level Image Features

Matthew Jenkins & Ian Reid

Much of the early Human Motion Tracking (HMT) work relied on traditional model-driven tracking techniques to predict the body location between frames, and thus limit the image search region over which to perform measurements. Such work sits firmly in the camp that would traditionally be labelled “tracking”. The problem with motion-model driven tracking of human body poses is that the motion of humans is often complex (i.e. difficult to model) and unpredictable (i.e. the motion model changes). Some work has tried to improve on this by learning motion models from real human motion data. However, this approach relies on having seen training examples of the motion being observed (making it very sensitive to the particular training data).

In contrast, a number of researchers (e.g. Viola, Jones & Snow) have developed methods of locating people in image scenes use machine learning algorithms such as AdaBoost or Support Vector Machines (SVM) which are trained beforehand. The emphasis in this work is on “detection” and usually restricted to pedestrian detection because of the inherent difficulties associated with training a system to cope with the wide variety of common human poses. Further, no pose is returned by these algorithms, just an overall body location, making them unsuitable for many applications of HMT. There are some attempts (e.g. by Agarwal & Triggs) to use machine learning techniques to learn the mapping from 2D image features to 3D pose. Impresssive results have been demonstrated (mainly on fronto-parallel walking motion), but again it is not certain whether it is possible to compile suitable training sets for more complex and unknown motions.

A sensible half-way house might be body-part detection incorporated in a more traditional tracking framework. Belief Propagation (BP) is a popular method for combining body-part detections (see, e.g., work by Sigal et al.) allowing sampling directly from image-based part proposals and combining these in a probabilistically correct manner. However, BP is computationally very expensive (especially in a non-parametric framework), and the output is not the desired joint posterior over all body parts, but rather the marginal distributions of each body part. This necessitates a post-processing step (adding to the computational expense and complexity). Bowden et al. use RANSAC and a strong learned prior pose model to combine the parts. This suits their application area (HCI) well, as the subjects are likely to have reasonably restricted poses, but applying the technique to more complex motions and poses may raise problems in achieving suitable training data. Further, part-based methods rely on high-level detectors correctly locating the body parts in each frame (often especially relying on the face). This is certainly not always the case, leading to tracking failure in these situations.

Similarly, our approach also combines the benefits of top-down model-driven tracking techniques (e.g. temporal consistency, use of a motion-model to track through occlusions) with the robustness of bottom-up data-driven approaches (e.g. recovery from tracking failure using global image information, no motion model so can handle complex motions). In contrast to the work above, we deliberately avoid the use of detection which requires a large degree of training (such as AdaBoost). Instead, we use persistent low-level features, tracked from frame to frame, to guide our sampling of the pose-space. We believe that using low-level image features to drive our tracking is desirable since: (i) they are efficient to detect (so direct image-driven tracking is possible at low computational expense); (ii) they require no a priori knowledge of the appearance of the person being tracked; and (iii) while each individual feature detection does not contain much pose information,
combining the information from multiple persistent low-level features (of different types) gives significant information on the pose of the person (or at least the motion undergone by the person).

We achieve the synthesis of top-down and bottom-up approaches using importance sampling techniques. We approximate the optimal strategy identified by various authors, e.g. Doucet (equation 1), which suggests modifying our motion-model predictions using the current image measurements, $z_k$, and sample from that function. This extension is rarely used for HMT, as full human body poses are difficult to sample directly from the image (with many techniques essentially attempting $\sim 30$-dimensional state-space search using Condensation).

$g(x_k^{(i)}|z_0:k, x_{0:k-1}^{(i)}) = p(x_k^{(i)}|z_k, x_{k-1}^{(i)})$  \hspace{1cm} (1)

We have implemented a basic low-level feature tracker which detects Harris corners in a frame and associates them with limbs using our current body pose estimate. We use weak motion model, uncertainty-gating and feature-descriptor matching to determine their motion over time. Only features which persist over a few frames are used in the next stage of the human body tracker.

The low-level feature tracking produces a large number of constraints on the human body motion. Rather than using the low-level image features to fix the location of a particular limb part, we ‘modify’ our motion-prior to take the measurement into account. This allows us to sample from an importance function which is similar to the optimal one (equation 1). We use constrained nonlinear optimisation of the motion prior to achieve this. Thus we can constrain a particular point on the arm (say) to appear at a particular image location. Parameters of the sampled states will be as close as possible to the motion-prior whilst remaining consistent with the constraint. Further, the uncertainty on the low-level tracking can be incorporated into the constraint. By sampling from the prior constrained by individual low-level features we obtain a large number of partially correct proposals, which can be weighted using the correct importance weights $w_k^{(i)} = \frac{p(z_k|x_k^{(i)})p(x_k^{(i)}|x_{k-1}^{(i)})}{g(x_k^{(i)}|z_0:k, x_{0:k-1}^{(i)})}$, with $g(\cdot|\cdot)$ chosen to represent the distribution from which the samples are drawn.

The samples obtained from the various low-level image features are combined using the crossover operator from Genetic Algorithms, to produce a sample from the fully measurement-constrained prior function. Previously Deutscher & Reid used the crossover operator for mode-finding applications, but we implement rigorous importance sampling so that the correct posterior is computed. We have had to determine the proposal density which is appropriate for samples formed using the crossover operator. The Constrained Prior sampling strategy approximates sampling from $p(x_k^{(i)}|z_k^{(f)}, x_{k-1}^{(i)})$ (where $z_k^{(f)}$ is the measurement of a particular low-level feature), and so the crossover operator allows us to sample from something like $p(x_k^{(i)}|z_k, x_{k-1}^{(i)})$: the optimal proposal density.

The system we have presented uses full-frame low-level feature detections to guide a model-driven high-level person tracker. It does not rely on a priori knowledge of the appearance or likely pose variation of the person being tracked nor on inaccurate motion models. Instead image cues are used to direct the tracking process, uniting traditional model-driven tracking and modern image-driven techniques.