VIGOUR: A System for Tracking and Recognition of Multiple People and their Activities

Jamie Sherrah and Shaogang Gong
Department of Computer Science, Queen Mary and Westfield College, London E1 4NS, UK
[jamie|sgg]@dcs.qmw.ac.uk

Abstract

Tracking multiple people and their behaviours is a fundamental task for visually mediated interaction. We present VIGOUR, a platform for simultaneously tracking of multiple people and recognition of their behaviours for high-level interpretation. Through perceptual integration, different types of visual information are fused to combine the benefits of different techniques. Robust low-level visual cues such as skin colour and motion are used to focus attention and facilitate real-time tracking. VIGOUR tracks behaviours using gestures and head pose to produce a high-level behaviour representation for subsequent interpretation. The system is able to track three people and recognise their gestures simultaneously in real time.

1. Introduction

Modeling of human behaviour is becoming an increasingly important and active area of research in computer vision. Several approaches exist, from full 3D modelling of the human body parts [1, 4] to statistical modelling of patterns from 2D views [3, 6]. However, under the constraint of real time processing, and given limited-resolution imagery and commonly-available hardware, many of these approaches are infeasible. We adopt a philosophy of integrating multiple complementary, inexpensive vision modules for real time tracking of human subjects and modelling of their behaviour.

Theories for modular perception in both artificial and biological systems have long been held [10]. Systems consisting of several integrated sub-systems benefit from the strengths of each system, can have redundancy and fault tolerance, and provide increased accuracy and robustness through averaging. A further benefit of composite computer vision systems is that different modules may be based on different assumptions. When the assumptions of one module are violated and it becomes unreliable, the rest of the system can still function. In this way the assumptions and constraints can be “factored out” [2]. The difficulty lies in fusing the available data sources in such a way as to exploit the possible benefits (improved robustness, redundancy) and avoid the pitfalls (strong inter-dependency, wasted computation). In this work an integrated system is presented that tracks multiple people and recognises their behaviours (head turning and gestures).

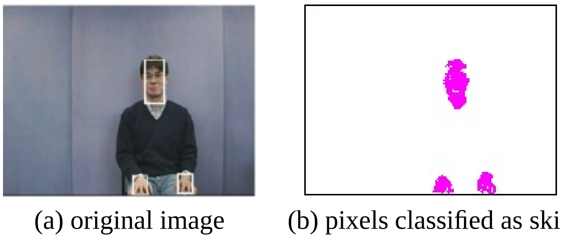
2. VIGOUR

An integrated *Visual Interface for Gestures and behaviOUR* (VIGOUR) was designed as a platform for investigating visually mediated interaction (VMI) methodologies. The current system uses a single pan-tilt-zoom camera as its only input, and integrates the following perceptual modules: (1) pixel-wise motion from frame differencing, (2) pixel-wise skin colour classification [8], (3) clustering into potential regions of interest, (4) support vector machine (SVM) for face detection [5], (5) person tracker to track head and hands, (6) gesture recognition [7], and (7) head pose estimation using similarity-to-prototypes [9]. In order to operate in real time, it is essential that the system focus on areas of interest and ignore irrelevant regions. Also, not every module need be active at each time instant. Some modules are used only to bootstrap the system into normal running mode. In the next section, the process of bootstrapping to find the people in the scene is described. The subsequent sections then describe the process of tracking the people and their behaviour.

2.1. Finding People

Initialisation or “bootstrapping” of any vision system is one of the most fragile and glossed-over steps. VIGOUR achieves robust bootstrapping through a trade-off of computational expense for accuracy. An SVM is used to detect near-frontal faces in regions of interest, which then initialises the person-tracking system. In general, the SVM would be too expensive to employ for each frame. Smaller regions of attention must be identified before searching for the face. Skin colour clusters are appropriate regions of interest in this case. The primary assumption made by the system is that the subjects are initially facing the camera and their faces are un-occluded.

The first step of bootstrapping is the calculation of a skin colour probability for each pixel in the image. Probabilities come from a mixture-of-Gaussian probability model in hue-saturation space. The model is trained beforehand from example pixels using the Expectation-Maximisation algorithm [8]. Both a foreground and a background model are used and combined using Bayes' rule. The image is subsampled in this step to reduce computational expense. The skin colour probabilities are thresholded to give a binary image of skin/non-skin pixels. An example of this output is shown in Figure 1.



(a) original image (b) pixels classified as skin

Figure 1. Example of skin colour detection.

In the next step, the skin colour image is passed to a clustering module to define skin blobs in the image. The output of the clustering module is a set of rectangular regions of skin colour (see Figure 1(a)), which could be faces, hands, or skin-coloured furniture or clothing. A multi-scale SVM scans inside these boxes at multiple positions to find faces [5]. When a face is found, it initialises a person tracker. An example of face detection is shown in Figure 2. The important point is that once the expected number of people in the scene has been found, the relatively expensive skin colour, clustering and SVM modules are deactivated. In this way, the data fusion method does not suffer from unnecessary computation.

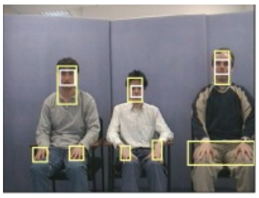
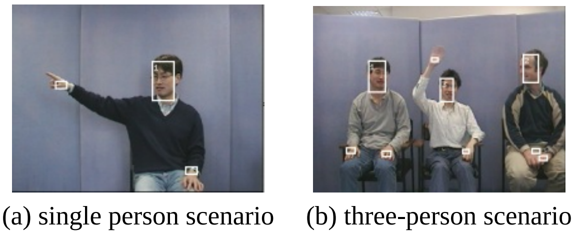


Figure 2. Example of SVM face detection. All outer boxes are skin clusters, while inner boxes are detected faces.

2.2. Tracking People

Given an image region known to be a frontal face, the task of the person tracker is to track the subject's head and hands. We use an inexpensive but also less robust technique for speed. We define a generic box tracker that tracks a box of possibly-moving skin. The person tracker consists of one box tracker for the head and one for each of the hands, initialised with different parameters. The head box tracker is initialised using the SVM face detection. The hand box trackers are initialised heuristically with respect to the head position. The box trackers have fixed size and track position only. An example of the person tracker is shown in Figure 3.



(a) single person scenario (b) three-person scenario

Figure 3. Examples of the person tracker.

Each box tracker operates under a fusion of adaptive skin colour and motion detection. Given the previous position of the box, skin colour and motion (frame differencing) are computed in a sub-sampled region within and surrounding the box. Skin colour is computed using the box's own adaptive colour model [8]. This allows the box tracker to adapt to individuals under changing lighting conditions due to movement, rather than remaining with the global skin colour model that must accommodate invalid skin colours for that individual. The pixels within the search region are then used to form a centroid (c_x, c_y) for the position of the box at the current time frame. Note that by calculating motion and skin colour only in the region of the box being tracked, excessive computation is avoided. By using motion as an additional but independent source of information, fault tolerance results. The person tracker is, however, directly dependent on the skin colour model, which can lose track by adapting to background regions.

2.3. Recognising Gestures

In some situations, a subject may need to highlight the focus of attention using emphatic gestures such as pointing and waving. Our previous method for recognising gestures used a spatio-temporal trajectory of global motion-based features as a representation [7]. Here we employ similar modelling and matching methods, extended to use a different object-centred representation. Gestures are recognised using a statistical matching approach. Features are extracted from each frame in a sequence, and the feature vectors concatenated to form a spatio-temporal trajectory. The features from frame i are collected into a feature vector $\mathbf{z}_{t_i} = [f_1, \ldots, f_d]^{\mathrm{T}}$, where t_i is the time at which frame i was captured, and d is the number of features. The resulting feature vectors are temporally ordered and concatenated to result in a trajectory of features $\mathbf{z} = \{\mathbf{z}_{t_1}, \ldots, \mathbf{z}_{t_n}\}$, where n is the number of frames in the sequence, and $t_n - t_1$ is the duration of the sequence.

The trajectories form structures in spatio-temporal feature space that can be modelled and discriminated. Some examples are shown in Figure 4. In order to perform recognition, gesture models are generated from a database of training sequences, and a Maximum Likelihood technique is used to classify novel gestures. The matching algorithm linearly scales the novel trajectory over time and matches backwards in time to the model using a Gaussian matching function. The features used determine the robustness and invariance of the method. In [7], it was assumed that only a single person was in the field of view at one time, and low-order moments from the whole image were used as features. We refer to this as the global method. Here we allow for several people in the field of view using a person-centred approach. The features used are the x and y offsets of the hands relative to the head box. This makes the representation translation-invariant. The features are scaled by dividing by the head box width. This gives the method scale tolerance.

The assumptions made by the gesture recognition module are implicitly controlled through the training set. In our case, we assume that the subject is seated and facing the camera. The output of this system is a set of likelihoods, one for each gesture being modelled. Figure 5 shows an example of the gesture likelihoods for a test sequence in which the subject pointed right, then waved, then pointed left.

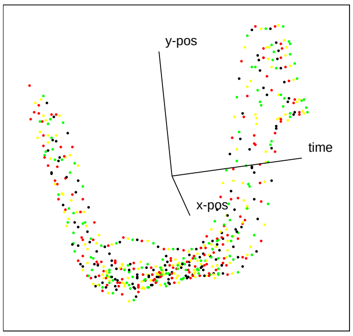


Figure 4. Example of spatio-temporal trajectories formed by an ensemble of waving gestures. Each point represents the position of left hand relative to head for one frame.

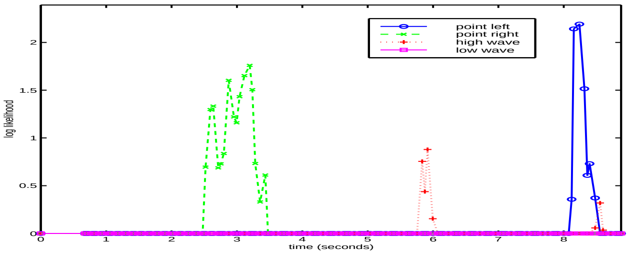


Figure 5. An example of gesture likelihoods for pointing and waving gestures. The sequence contains a point right, then a wave, then a point left gesture.

3. Examples of VIGOUR

We now present two examples of the working VIGOUR system. In the first example, Figure 6, it is assumed that there is a single person in the field of view. The subject is tracked and the camera responds to gestures in real time by panning and zooming. The camera response is heuristically determined, which is appropriate given that there are few possibilities for interesting areas in the scene. In the first instance (Figure 6 (a)) the subject waves, and the camera responds by zooming in for a closer view of the subject's head. In the second instance (Figure 6 (b)), the subject points to his right in order to draw attention to a colleague out of the current view. The camera responds by panning to accommodate the second subject in the view.

In the second example the camera is assumed fixed with a wide-angle view, and there are three people in the scene. VIGOUR tracks three subjects simultaneously and recognises their gestures and estimates their head pose. The subjects are labelled A, B and C from left to right. In the scenario, C waves to gain attention and begins speaking, then he points to A. Annotated example frames from the scene are shown in Figure 7. The tracked bounding boxes around the head and hands of each person are shown, along with the tracked face position. The dial above the head shows the estimated head pose.

4. Discussion and Conclusion

A system for extracting information from dynamic visual scenes in the context of VMI has been presented. Several distinct perceptual modules are integrated to simultaneously track several subjects, recognise their gestures and estimate their head pose. The outputs of this system are fed to a higher-level interpretation model. It is unnecessary for all systems in VIGOUR to be thoroughly robust. For example, the inaccurate head pose of subject A in Figure 7

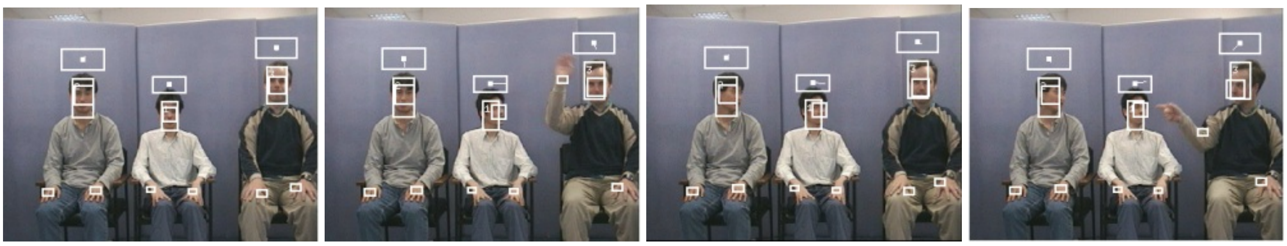
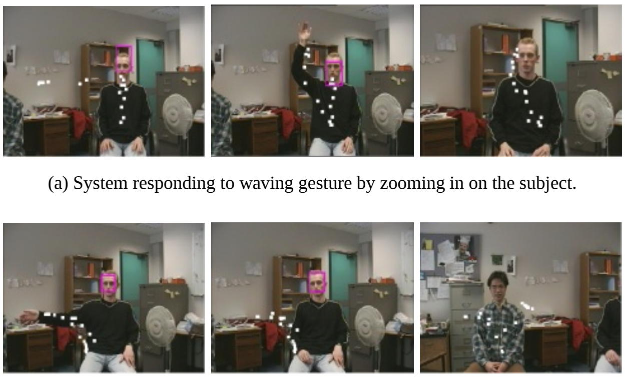


Figure 7. Example of the output of VIGOUR tracking three people simultaneously. For each person the top box shows the head pose using a dial, the centre boxes show the tracked head (outer) and face (inner), and the lower boxes show the tracked hands.



(a) System responding to waving gesture by zooming in on the subject.

(b) System responding to pointing gesture by panning to other subject.

Figure 6. Example of VIGOUR responding to gestures of a single person in the field of view in real-time. Each white square indicates the global centroid of the motion field for a single frame. These centroids were among the features used to recognise the gestures.

may be seen as irrelevant by a high-level interpretation system because C has waved to gain attention. In future work, the interpreting system will feed its knowledge back into VIGOUR to correct errors and guide focus of attention online.

An important question concerning VIGOUR is how to evaluate its performance. It is possible for almost all subsystems to be evaluated individually. For example, the detection rate of the SVM was evaluated from test sets. However, the tolerability of failure depends on how interdependent the modules are. If each sub-system has a probability of failure p_i, then the probability of at least one subsystem failing is $1-\prod_{i=1}^{N}(1-p_i)$, where N is the number of modules. For example, if each of our seven sub-systems has a probability of failure equal to 0.05, the overall probability of successful operation is 0.302, which is quite unacceptable. Clearly the system must be designed in such a way as to make the modules semi-independent of each other, such as the skin colour and motion modules. VIGOUR currently relies more on the skin colour and person tracking units, a dependence that could be undesirable.

References

[1] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. In *ICCV*, pages 624–630, Cambridge MA, 1995. IEEE.

[2] J. Clark and A. Yuille. *Data Fusion for Sensory Information Processing Systems*. Kluwer Academic Publishers, 1990.

[3] T. Darrell and A. Pentland. *Artificial Neural Networks with Applications in Speech and Vision*, chapter Recognition of Space-Time Gestures using a Distributed Representation. Chapman and Hall, London, 1993.

[4] D. M. Gavrila and L. S. Davis. Towards 3-d model-based tracking and recognition of human movement: a multi-view approach. In *Int. Conf. on Auto. Face and Gest. Recog.*, pages 272–277, Zurich, 1995.

[5] Y. Li, S. Gong, and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. In *Int. Conf. on Auto. Face and Gest. Recog.*, pages 300–305, Grenoble, France, Mar. 2000.

[6] J. Martin and J. L. Crowley. An appearance-based approach to gesture-recognition. In *Int'l Conf. on Image Analysis and Processing*, Florence, Italy, Sept. 1997.

[7] S. McKenna and S. Gong. Gesture recognition for visually mediated interaction using probabilistic event trajectories. In *BMVC*, volume 2, pages 498–507, Southampton, England, Sept. 1998.

[8] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using colour. In *Int. Conf. on Auto. Face and Gest. Recog.*, pages 228–233, Nara, Japan, 1998.

[9] J. Sherrah and S. Gong. Fusion of perceptual cues using covariance estimation. In *BMVC*, volume 2, pages 564–573, Nottingham, UK, Sept. 1999.

[10] S. Ullman. Visual routines. *Cognition*, 18:97–159, 1984.