Mathematical modelling of animate and intentional motion

Jens Rittscher^{1*}, Andrew Blake², Anthony Hoogs¹ and Gees Stein¹

¹GE Global Research, One Research Circle, Niskayuna NY 12309, USA

(anthony.hoogs@research.ge.com, gees.stein@crd.ge.com)

²Microsoft Research, 7 Jf Thomson Avenue, Cambridge CB3 0FB, UK (ablake@microsoft.com)

Our aim is to enable a machine to observe and interpret the behaviour of others. Mathematical models are employed to describe certain biological motions. The main challenge is to design models that are both tractable and meaningful. In the first part we will describe how computer vision techniques, in particular visual tracking, can be applied to recognize a small vocabulary of human actions in a constrained scenario. Mainly the problems of viewpoint and scale invariance need to be overcome to formalize a general framework. Hence the second part of the article is devoted to the question whether a particular human action should be captured in a single complex model or whether it is more promising to make extensive use of semantic knowledge and a collection of low-level models that encode certain motion primitives. Scene context plays a crucial role if we intend to give a higher-level interpretation rather than a low-level physical description of the observed motion. A semantic knowledge base is used to establish the scene context. This approach consists of three main components: visual analysis, the mapping from vision to language and the search of the semantic database. A small number of robust visual detectors is used to generate a higher-level description of the scene. The approach together with a number of results is presented in the third part of this article.

Keywords: computer vision; event recognition; visual tracking; video annotation; mapping from vision to language

1. INTRODUCTION

Rather than understanding the perception of biological motion we intend to construct a machine that is able to recognize certain biological motions. The crucial question before we derive any mathematical model is the choice of which representation to use. We are particularly interested in computationally efficient approaches that facilitate the instantaneous classification of the observed motion. Moving light displays (Johansson 1976) are often used to test our ability to recognize biological motion. The nature of the internal representations the human visual system uses to recognize these sequences is of particular interest to this work. Bülthoff & Bülthoff (2003) designed a number of experiments to determine whether or not the human recognition capability depends on the viewpoint of the scene. Based on the results of these experiments they conclude that the subjects' recognition performance is strongly tied to the familiarity of the viewpoint from which the sequences are captured. The fact that stereo-depth information only marginally improves the recognition performance furthermore indicates that the internal motion models emphasize the 2D trajectory structure of the sequence over its 3D structure.

Hence the guiding principle of this work is to find an

appearance-based approach such that it is not necessary to estimate the person's pose in order to classify the type of motion. Therefore we avoid the use of an articulated model such as a stick figure model or a 3D skeleton. The pose estimation itself is central to the problem of human motion capture. Recent advances in computer vision (Deutscher *et al.* 2000; Sidenbladh *et al.* 2000) allow us to solve the pose estimation problem without having any kind of markers attached to their bodies. But in order to resolve ambiguities multiple views are necessary to accurately estimate the pose. In turn we intend to recognize the type of motion based on the apparent motion. In the current context of this article, one interesting point of discussion would be if the appearance-based information alone is sufficient for movement generation as discussed by Schaal *et al.* (2003).

Two distinctly different approaches to modelling biological motion are presented in this article. Following the principles of pattern theory (Grenander 1976–1981), or 'analysis by synthesis', the first approach in § 2 uses a generative model to recognize the observed biological motion. Here, the apparent contour of a person is described by an active contour (Kass *et al.* 1987). The deformation of the contour is controlled by a low-dimensional configuration space. The dynamics of the apparent contour is explicitly modelled as a stochastic process. We demonstrate that finely tuned stochastic processes can be used for both the anticipation and the classification of the observed motion.

The type of biological motion is recognized based on the dynamical information alone. No information about

* Author for correspondence (rittische@crd.ge.com).

One contribution of 15 to a Theme Issue 'Decoding, imitating and influencing the actions of others: the mechanisms of social interaction'.

Figure 1. Role of context information. In each of these examples the scene context helps to disambiguate the type of motion. One should notice that there are often very fine and subtle distinctions which allow us to recognize the type of motion.

the shape of the apparent contour is used. It has been shown elsewhere that the shape information can be used effectively to detect people and estimate their pose. To facilitate the detection of pedestrians from a moving car Gavrilu and Philomin (Gavrilu & Philomin 1999; Gavrilu 2000) represented a large collection of shapes using a hierarchy of templates that are based on edge-maps. The strength of this approach is that it captures the variety of shapes very efficiently. The similarity of two shapes is computed using a metric that is based on the distance transform (Borgefors 1988). The online matching is realized as a simultaneous coarse-to-fine search over the shape hierarchy and transformation parameters. The hierarchy itself is learnt from training examples using a stochastic optimization technique. Although the main application area of this system is the detection of pedestrians, one can now replace the shape contour with a large collection of example shapes in order to track the object. Toyama & Blake (2001) propose a *metric mixture* model which combines exemplars in a metric space with a probabilistic treatment.

A related but different approach is to recognize the type of motion directly from the spatio-temporal features of the image sequence. In § 3 we explain how the spatio-temporal structure of human motion can be encoded in a very low-dimensional vector. Standard pattern recognition techniques are used to classify sequences containing a variety of different motion classes. Using this technique it is no longer necessary to localize the apparent contour. Consequently, the task of tracking the person is simplified.

If we intend to give a higher-level interpretation rather than a low-level physical description of the observed motion the scene context plays a crucial role. For example, in all the images displayed in figure 1 men raise their arms, but they perform very different activities. This also relates to the question of which biological motions a machine could possibly recognize. Regardless of which feature space and representation is used none of these approaches measure up to the human ability to learn and recognize biological motion. The task of recognizing a simple motion like walking, independently from scale and viewing angle, is an enormously challenging problem. Apart from

the ability to recognize canonical biological motions like walking and running regardless of viewpoint, scale and occlusion, humans are also able to generalize concepts and make very fine context-dependent distinctions. For example, once a child is aware of the difference between walking and running it has no trouble making the analogous distinction for animals (Bloom 2000). Whereas the generalization itself is arguably a very difficult problem, the importance of contextual information is much more apparent.

Once we accept that contextual information is helpful or even crucial, the question of how such knowledge should be represented arises. Such knowledge can be learnt from examples. This particular problem is difficult as we are interested in a very wide variety of scenes. A typical daily news broadcast on television serves as a good example. Hence the challenge of selecting models for motion recognition is to construct a model which is sufficiently abstract that it allows us to recognize events which are different but still related to instances in the training set. A common approach to this problem is the use of probabilistic networks or graphical models (Pearl 1988). Although the field of machine learning is far advanced there are of course limits of the current ability to learn models of very complex systems. An alternative to this approach is the use of a semantic knowledge base. Prior to searching the semantic knowledge base the video is annotated by a set of simple visual descriptors. This information is used to establish the scene context. This context information is then used to give a more specific description of the image sequence. This complex system consists of three different components: the visual analysis, the mapping from vision to language and the search of the semantic database. A description of these components together with a number of results is presented in § 4.

2. DYNAMICAL MODELS

The approach of tracking people (Baumberg & Hogg 1994; Blake & Isard 1998) using their apparent contour is mathematically a very rich model which extends naturally to the recognition of certain periodic biological

Figure 2. Discrete eigenvalues of motion models learnt from three different walks. Three different walking sequences of one human are tracked. For each sequence a second order ARP is learnt using the maximum-likelihood learning rule. Only the periods in which the walker is in a steady state are used to learn the model. It is crucial to observe that only the first four modes (displayed as black circles) are similar for all three models. From these only the first two nodes are significant since they have a time constant of $\beta^{-1} > 1$ s. All remaining modes have a damping constant of $\beta^{-1} < 0.2$ s.

motions. The apparent contour of an object is modelled as a spline contour. The deformation of this contour with respect to a template is controlled by a low-dimensional linear state space, the shape space. The objective of the visual tracking system is the estimation of the state vector X_t based on a set of measurements z_t taken from the current frame of the image sequence at time t . A stochastic filter (Gelb 1979; Lütkepohl 1993) is then employed to estimate the state X_t given the history of the measurements $Z_t = \{z_0, \dots, z_t\}$. A learnt dynamical model is used as a predictor. Under certain assumptions the Kalman filter (Gelb 1979) is the optimal linear filter to estimate the state X_t at time t .

But rather than having a single estimate as in the Kalman filter here the posterior probability $p(X_t|Z_t)$ is propagated over time. The probability density $p(X_t|Z_t)$ is represented in a non-parametric form by a set of K samples $\{X_t^{(i)}\}_i$. Each sample $X_t^{(i)}$ has a likelihood weight π_i associated to it, such that $\sum_i \pi_i = 1$. The interpretation of such a particle set is that if the set is *resampled*, meaning that an X is chosen to be one of $X_t^{(n)}$, with probability proportional to its weight $\pi_i^{(n)}$, that X is distributed (approximately) according to the posterior p_t . A condensation or particle filter (Gordon *et al.* 1993; Isard & Blake 1996; Kitagawa 1996) is applied to propagate $p(X_t|Z_t)$ over time. In principle this framework is a generalization of the Kalman filter which also allows the use of nonlinear dynamics in the form of multi-class dynamics.

(a) Multi-class dynamics

Multi-class dynamics are represented by appending to the continuous state vector x_t a discrete state component y_t to make a 'mixed' state $X_t = (x_t, y_t)^T$, where $y_t \in \{1, \dots, Y\}$ is the discrete component of the state, labelling the class of motion. Corresponding to each state $y_t = y$ there is a dynamical model, taken to be a Markov model of order K that specifies $p_t(x_t|x_{t-1} \dots x_{t-K})$. Each class y has a set (A^y, B^y, d^y) of dynamical parameters, which are learnt from example trajectories (North *et al.* 2000). In addition, and independently, state transitions are governed by the transition matrix for a first-order Markov chain:

$$P(y_t = y' | y_{t-1} = y) = M(y, y').$$

As mentioned earlier, the stochastic process used in the prediction step is the underlying model for the dynamics of the spline contour. To understand why this framework is suited to model certain periodic biological motions a more detailed knowledge of the dynamical process is necessary. A linear-Gaussian Markov model of order K is an ARP (Lütkepohl 1993) defined by

$$x_t \sum_{k=1}^K A_k x_{t-k} + d + B w_t \quad (2.1)$$

in which each w_t is a vector of N_x independent random $N(0,1)$ variables and $w_t, w_{t'}$ are independent for $t \neq t'$. Note that the stochastic parameter B is a first-class part of a dynamical model, representing the degree and the shape of uncertainty in motion, allowing the representation of an entire distribution of possible motions for each state y . In continuous time a second-order Markov process is governed by a stochastic differential equation, the Fokker-Planck or Kolmogorov equation (Karatzas & Shreve 1991)

$$m\ddot{x} + c\dot{x} + kx = bw,$$

where w is white noise. When there is no noise present, i.e. $w = 0$, this is an oscillator with mass m , damping constant c and stiffness k . The sample path of such a process in continuous time in one dimension is a damped harmonic oscillation. The damping rate is denoted by β . A multivariate process can be decomposed into damped harmonic oscillations along different directions of the configuration space. An example of a modal analysis of a multivariate process is shown in figure 2. Depending on the magnitude of the eigenvalue λ each mode can be characterized as unstable, exponentially decaying or harmonic. Harmonic modes with a high damping rate influence the motion only for a very short amount of time, and the least damped modes characterize the motion.

Our experiments indicate that ARPs used for the anticipation of motion (linear, stochastic differential equations and their discrete embodiment as ARPs) are capable of modelling certain repetitive human motions such as running and walking. For example, the modal analysis of three different learnt models for walking presented in figure 2 indicates that a significant number of modes have a

Figure 3. Motion classes used in the experiment. In order to illustrate the motion, contours of previous time-steps are superimposed on one frame. (a) Half star: a ‘star jump’ without arm moment. Notice that both motions begin with an upwards acceleration. Hence it is difficult to discriminate between both motions. (b) Pure jump, i.e. jumping up and down without lateral arm or leg movement.

Figure 4. Examples of the different segmentations. The motion class pure jump is shown in red, half star in blue. The top row corresponds to the ground truth obtained by hand segmenting the sequence. A two-state ‘condensation’ filter without importance sampling is used with $n = 2000$ particles. The third row displays the segmentation obtained using partial importance sampling on the discrete state with $n = 400$ samples. Note that the quality of the segmentations does not get worse when the sample size reduces. The crucial role of the transition matrix is documented in the bottom row. Here, a transition matrix with $M(y, y) = 0.9$ and importance weights with $g(y, y) = 0.8$ are used; hence the expected duration of the motion is no longer correct.

damping constant β^{-1} of less than 0.2 s. They are ‘unused’—i.e. discarded by the learning algorithm. Hence there is no need to employ a more complex stochastic process to model the type of motion.

By introducing the class of motion as a discrete variable of the state X_t , the particle set which represents the posterior distribution $p(x_t|Z_t)$ enables us not only to track the apparent contour but also to estimate the class of motion at any given time t . Hence we have developed a system where classification feeds back into the perception of

motion which means perception and classification are inextricably bound together. In this approach the motion is classified instantaneously and does, for example, in a surveillance application provide the ability to raise an alarm. In case no instantaneous classification is necessary the classification results can be improved by smoothing as shown in North *et al.* (2000). The following experiment demonstrates that an image sequence can now automatically be segmented into subsequences that contain only one type of motion.

Figure 5. Misclassification graphs. The average error of three tests for each of the three filters on a particular test sequence is displayed for different particle sizes N . Circles and dashed line, mixed state condensation; triangles and solid line, partial importance sampling.

(b) Experimental results

Two different motion classes were used for a classification experiment. A *pure jump*, i.e. jumping up and down without lateral arm or leg movement, and a *half star*, which is a 'star jump' without arm movement. In order to illustrate the two motions the contours of the previous time-frames are superimposed onto one frame (figure 3). A separate set of training sequences that contain only one type of motion was used to learn the motion models. In all the experiments the tracking process is initialized by hand. In order to get a very good model the maximum-likelihood learning rule was used on a sequence of 8 s in length. Note that a combination of these two motions illustrates both types of classification problems: the lateral leg movement makes it easy to decide when the person is in motion class *half star*. The jumping up from the bending down position is, however, very difficult. As can be seen in figures 4 and 5 we achieve an asymptotic misclassification performance of 10%.

We have demonstrated that a particle filter with mixed states can be used for classifying motions online. In the approach presented by Bregler (1997) the discrete states are used up for modelling atomic motions. In that context a bi-level recursive algorithm (Rabiner & Bing-Hwang 1993) could be tried for automatic segmentation. But these algorithms are notoriously computationally expensive. Here, we show that a single autoregressive model is a serious candidate as a model of atomic motions. This leaves the discrete state free for classification. A Markov chain is used to model the state transitions and to apply long-term continuity constraints. It was demonstrated that the Markov chain can be applied more effectively by using partial importance sampling. In general, importance sampling (Hammersley & Handscomb 1964) is a method for variance reduction. The details of this method can be found in Rittscher & Blake (1999). Learning the motion models from a mixed sequence as in North *et al.* (2000) could potentially reduce the asymptotic error rate of 10%. It could well be possible that human intuition imposes a

Figure 6. Spatio-temporal surface. A sequence of mixed aerobics exercises (as shown in figure 10) is used to compute a space-time surface. The surface is extracted from the reprocessed sequence using an implementation of the 'marching cubes algorithm' (Schroeder *et al.* 1992). Note that in the latter part of the sequence the arm motion is clearly visible.

segmentation which is sub-optimal with respect to ARPs. Both motions have a number of similarities so it might be necessary to work with more than two atomic motion models. The motion classes used in this experiment could then correspond to the occurrence of one or two atomic motion models.

But there is one other concern. Tracking complex motions of humans using a contour tracker is very labour intensive. It is especially difficult to initialize the visual tracking process. Although the background of the scene is very simple we failed to track the full star jump or jumping jack (as shown in figure 10a(i)). Furthermore, there is effectively no variation of pose in these sequences. This raises the question whether it would be more desirable to develop a more direct approach to motion classification. If one could classify the class of motion using the raw image data it would be no longer necessary to rely on a contour tracker. Such an approach might also solve the initialization problem for contour tracking. The development of such a method will be discussed in the following section.

3. CLASSIFYING MOTIONS DIRECTLY FROM SPATIO-TEMPORAL FEATURES

As just mentioned, the aim is now to recognize biological motion directly from the spatio-temporal features of the image sequence. Black & Jepson (1998) use the optical flow field of the entire image to recognize gestures and expressions. One drawback of this method is that the flow field is not very well localized and can easily be corrupted by background motion. It turns out that representing the entire image sequence as a space-time cube or *XYT* cube, as opposed to analysing adjacent frames, is more appropriate for motion analysis. Niyogi & Adelson (1994a,b), for example, fit a spatio-temporal surface rather similar to the one shown in figure 6 to estimate the parameters of a stick-figure model. They use these parameters to recognize people by their gait. But as motivated in the introduction,

our thesis is that motion recognition should not require any articulated model.

Without making any use of the underlying structure of the motion pattern Zelnik-Manor & Irani (2001) present a non-parametric approach. The imagery is analysed at three or four temporal scales and normalized spatial-temporal gradients at each scale are computed as

$$\frac{\nabla I(x, y, t)}{|\nabla I(x, y, t)|}, \quad (3.1)$$

where image sequence is denoted by I . Hence at each of the L scales one gets a 3D feature vector on each voxel location. To reduce the computational complexity, they avoid estimating the joint density and treat each feature component and scale independently. The comparison of these feature histograms is very difficult because these histograms are typically dominated by feature points for which $|\nabla I|$ is almost zero. This is why they formulate a distance metric that uses a weighting factor for each histogram bin, i.e.

$$D(h, h') = \sum_i w_i (h(i) - h'(i))^2, \quad \text{where} \quad w_i = (h(i) - h'(i))^{-1}. \quad (3.2)$$

In the case that a number of sequences for each event type are available, they refine this distance measure by computing the mean and variance in each histogram bin and essentially computing cluster centres, each representing a class of motion. As mentioned before, points for which $|\nabla S|$ is almost zero swamp the statistic. This effectively causes a large amount of observation noise. As can be observed in figure 8, the salient information is contained in the motion boundaries. This is why a conservatively set threshold can be applied to eliminate such data points from the statistic.

Motions like running, walking and skipping can be characterized by the different intrinsic velocities of leg movement. The slices of the XYT cube exhibit a braided pattern which characterizes each type of motion. Cipolla & Yamamoto (1990) use these slides for the stereoscopic tracking of bodies in motion and introduce the term epipolar slices. Because the epipolar slice is an entity in space time, the braided pattern is directly related to the velocity profile of the motion. The braided pattern of leg motion, for example, consists of two self-intersecting curves, one for each leg. Hence, the velocity of the leg can be computed by estimating the outer normal to the curve. This is also illustrated in figure 7. The angle α of the outer normal can be computed as

$$\tan(\alpha) = \frac{\delta_t I}{\delta_x I}.$$

For practical purposes $\delta_x I$ and $\delta_t I$ are computed from the bandpass-filtered data. And in order to suppress feature points that are not associated with the motion boundary α is only calculated where $|\nabla I|$ exceeds a threshold, i.e. $|\nabla I(x, t)| > C$, as

$$\tan(\alpha(x, t)) = \frac{(I * \delta_x \varphi)(x, t)}{(I * \delta_t \varphi)(x, t)}. \quad (3.3)$$

The threshold C can be set conservatively. Its only purpose is to prevent the distribution $p(|\alpha|)$ from being

swamped by α -values that correspond to locations (x, t) at which the modulus of the gradient $|\nabla I(x, t)|$ is near zero. Because the $\tan$ function is antisymmetric, i.e. $-\tan^{-1}(-x) = \tan^{-1}(x)$, for every time t the modulus of α , $|\alpha|$, estimated from either curve, is identical. Hence the distribution of $|\alpha|$ will describe the velocity distribution of the motion pattern.

The distribution $p(|\alpha|)$ clearly depends on the position of the epipolar slice. Therefore a collection of slices will be used to characterize the motion pattern. The resulting feature vector consists of a set of skew factors $\gamma_1(I, y)$ computed from a slice at height y , i.e.

$$\gamma_1(I) := \{\gamma_1(I, y_1), \dots, \gamma_1(I, y_N)\}, \quad (3.4)$$

where N is the number of epipolar slices taken from the space-time cube. Hence we compress the image sequence I to a feature vector of length N . The estimate of the third moment or skewness, γ_1 (see, for example, Scheffe 1959), of the distributions is defined as

$$\gamma_1(\{\alpha_1, \dots, \alpha_N\}) = \frac{1}{N} \sum_{j=1}^N \left(\frac{\alpha_j - \bar{\alpha}}{\sigma} \right)^3. \quad (3.5)$$

The skewness of a distribution measures the degree of asymmetry. The measure does not depend on the location or scale (measured respectively by the mean $\bar{\alpha}$ and the variance σ). Hence a linear transformation of the distribution will not affect the skew factor γ_1 . For a symmetrical distribution the skewness, γ_1 , is evidently zero. A positive value of skewness signifies a distribution with an asymmetric tail extending to the right of the mean and vice versa. Three typical epipolar slices, ranging from fast to slow motions, are displayed in figure 8. Both the learnt distributions and the skewness, γ_1 , allow us to discriminate between the three different velocity profiles. We therefore conclude that it is sufficient to compute the skewness of the learnt distribution of a collection of epipolar slices and treat the vector of the skew factors as a feature vector. Results of this method are shown in figures 9 and 10. It should be noted that the vectors of skew factors shown in figure 9 give rise to a very natural interpretation of the characteristics of the three different motion types. This method orders these different walking styles in some kind of continuum where the two extremes are defined by running and walking. Here, we make use of the fact that the walking motions are periodic. In the next section we will demonstrate that the same method can be used to recognize non-repetitive motions when very short time intervals are used.

Many applications require the instantaneous classification of the type of motion. The skew vectors $\gamma_1(I)$ are computed for a set of consecutive frames or, in other words, of a spatio-temporal cube of a certain temporal length T . Apart from the noise of the measurement model itself, the distribution of the skew vectors $\gamma_1(I)$ for each class of motion will now depend on the length of the temporal window. This is due to the phase dependency of the motion itself. In a first experiment we attempt to classify each temporal window using a very basic model for the class densities. Like in the computation of the linear discriminant analysis in the previous section, the distributions of the skew vectors $\gamma_1(I)$ for each motion class are modelled as multivariate Gaussians with mean μ and

Figure 7. Idealized braided pattern. (a) An idealized braided pattern of a person walking under a front-to-parallel view. (b) The angle of the outer normal to one curve, α , measures the velocity of the leg at time t . The other curve is a mirror image of the first hence the angles of both other normals are related. This is because the tan function is antisymmetric, i.e. $-\tan^{-1}(-x) = \tan^{-1}(x)$. It is also clear that the angles of the inner normals correspond to those of the outer normals. The distribution of $|\alpha|$ hence characterizes the braided pattern. (c) A typical learnt distribution of $|\alpha|$ from an epipolar slice taken from the experimental data.

Figure 8. Skewness of the $|\alpha|$ distributions. This figure illustrates the effect of estimating the skewness $\gamma_1(6)$ of the learnt distributions for $|\alpha|$. (a) Examples of typical epipolar slices for (i) running, (ii) skipping and (iii) walking. The corresponding distributions of $|\alpha|$ are shown in (b). The reader should note that all three representations of the data—the raw data, the learnt distribution of $|\alpha|$ and the skewness of the distribution—allow us to discriminate the three different motion patterns. In these examples the foreground pattern (in black) is clearly visible. But as this method is based on the estimation of normals the representation does not depend on the intensity difference between foreground and background as long as it exceeds the threshold C (see equation (3.3)).

diagonal covariance Σ . The Mahalanobis distance between the skew vectors $\gamma_1(I)$, computed for each interval of length T , and the learnt centres μ_i is used to classify the observed motion.

We believe that it is important to test the use of the skew vectors $\gamma_1(I)$ using a very basic classification technique. Unlike HMMs, such an approach discards, of course, the temporal information about the mean duration of each motion. But parameters for models with hidden parameters are usually learnt using some form of expectation maximization learning rule. These algorithms do

not guarantee to find a local minimum and depend heavily on an initialization step. If the results obtained by a basic classifier are promising it will be possible to refine them using a more elaborate model.

We now benefit from the fact that it is no longer necessary to track the outline of the person in order to classify the type of motion. This allows us to test the newly developed method on examples that are considerably more complex than those used in § 2. The first dataset contains sequences of four different people running, skipping and walking. Owing to the low contrast between

Figure 9. Skew vectors $\gamma_1(I)$ for different walking styles. The skew vectors $\gamma_1(I)$ (3.5) are computed for a set of 20 epipolar slices. The exact position of the slices are indicated as red lines in (a). The skew vectors, $\gamma_1(I)$, for each of the sequences are presented in (b). The skew vectors are colour coded. Red corresponds to running, green to skipping, and blue to walking. It can be observed that the skew factors of top and bottom slices are implausible. This is a consequence of the naive threshold set in (see equation (3.3)). The amount of variation in the upper half of the body can be explained by the body motion with respect to the centroid of the body. As a result of computing the skew factors of the different epipolar slices, the different walking styles are ordered according to their average speed. In an experiment a linear discrimination analysis of the skew vectors $\gamma_1(I)$ for a set of four different people performing running, skipping and walking was used to demonstrate that the three different motion classes can be separated.

Figure 10. Skew vectors $\gamma_1(I)$ for the aerobics motions. In these experiments the skew factors $\gamma_1(I)$ (see equation (3.4)) are computed for a set of 40 epipolar slices. The exact position of the slices is again indicated as red lines in (i). The skew vectors $\gamma_1(I)$ for each of the test sequences are presented in (b). The skew vectors are colour coded. Red corresponds to jumping, green to a half-star jump and blue to the full star-jump. It should be noted that although this method is more suitable for analysing leg motion (see text) the presence of arm motion is detected correctly. Moreover the fact that the skew factors are a relatively smooth function with respect to y indicates that the estimates are not particularly noisy.

foreground and background it would be difficult to track the outline of the people using an edge-based contour tracker. The second set of test sequences consists of a set

of aerobics exercises similar to those used previously. But here we included the star jump as the third motion class. All sequences of people walking were recorded at the

Figure 11. Skew vectors $\gamma_1(I)$ for basketball player. The skew vectors $\gamma_1(I)$ (3.5) are computed for a set of 44 epipolar slices. The feature vectors are convolved with a smoothing filter to reduce the measurement noise (see text). As before, (a) shows the location of the epipolar slices. The distributions of the skew vectors of each motion class are modelled by a multivariate Gaussian with a diagonal covariance matrix. (b) The mean and standard deviation of each motion class. Four motion classes are used in this experiment: running (in red), walking (in green), turning (in magenta) and throwing (in blue). It can be easily observed that the throw is the only motion that involves a considerable amount of arm motion. The turning motion involves very little arm and leg motion. The motion classes running and walking, however, show some overlap.

same time and place. The group of people contained three males and one female all between 20 and 30 years old. In all, the dataset includes 77 sequences whose lengths vary between 2 s and 5 s. A simple blob tracker based on motion detectors was applied to locate the foreground window in every frame.

The second set of test sequences contains three different gymnastic exercises: jump, half-star jump and star-jump (or jumping jack) (see figure 10). These sequences do not require any preprocessing as the exercise is performed on the spot. Two sets of image sequences were analysed. The first set contains three sequences of a mixture of aerobics exercises shown in figure 10. The temporal length of the spatio-temporal cube T was chosen to be 25 fields or half a second. Like in the experiment shown in figure 10, a number of 40 epipolar slices were used to compute $\gamma_1(I)$. The distributions of the skew vector for each class, jumping, half-star and star-jump were learnt from a separate set of training sequences, each of which contained only one type of motion. In total, 114 half-second intervals were analysed. Out of these, eight were classified wrongly, which corresponds to an error rate of 7%.

These classification errors are due to measurement noise. To reduce this measurement noise we convolved each feature vector with a smoothing filter. This is justified by the following observations. First, we expect the true skew factors $\gamma_1(I_{3y})$, as mentioned before, to be continuous with respect to y . Second, the noise of the skew estimate on two adjacent slices $\gamma_1(I_{3y_i})$ and $\gamma_1(I_{3y_{i+1}})$ is assumed to be independent. One justification for this is that the

support of the filters, $\delta_x\varphi$ and $\delta_y\varphi$, needed to compute α (see equation (3.3)), on the two adjacent slices are disjoint. As a result the number of misclassifications is reduced to three out of 114 samples. This corresponds to an error rate of 2.6%.

Finally we analysed sequences of a person playing basketball. Two sequences, each of which is 50 s long, were used. Like in the previous example the length of the spatio-temporal cube T was set to be 25 fields or half a second. The skew vectors $\gamma_1(I)$ were again convolved with a smoothing kernel. The foreground region was tracked using a simple blob tracker. The training data were obtained by labelling each half-second interval in one of the sequences as belonging to one of the following motion classes: walking, running, turning and throwing. As before, each of these motion classes was modelled by a multivariate Gaussian with diagonal covariance matrix. The resulting means for each of the motion classes are shown in figure 11. As can be seen in the figure, the distributions of walking and running have some overlap. As a result, walking is occasionally misclassified as running. In total, 10 out of 99 samples are misclassified, which corresponds to a misclassification rate of 10%. All remaining classes of motion are identified correctly. Single frames of the sequence are shown in figure 12. Taking the difficulty of the sequence into account this is a very promising result. Potentially, this error rate results from the fact that during training some intervals that contain a transition from standing still to running were labelled as running. Assuming that the camera is stationary, this error rate can

Figure 12. An automatically annotated basketball sequence. Shown are single frames out of the automatically annotated sequence. The person runs towards the basket dribbling the ball. The ball is then thrown into the basket. After the ball is caught the person walks back. All of these stages are identified correctly. Only the decision between running and walking is sometimes ambiguous.

easily be reduced by taking prior information, such as the output from the blob tracker, into account. In comparison with the examples shown in the previous section, the task of visual tracking is a lot simpler. Here, a blob tracker like the one presented in Comaniciu *et al.* (2000) can be used to extract the motion trajectory of the foreground object.

4. SEMANTIC INFORMATION AND CONTEXT

The objective is now to give a higher-level interpretation of the imagery. As discussed in § 1, this is required to establish the context of the scene. We now present an approach that uses the information contained in a semantic database to establish this context information. The overall approach consists of three main components: visual analysis, the mapping from vision to language and the search in the semantic database. Before details of these components can be given, the nature of the semantic database used here needs to be described. The current capability of the system will be discussed in § 4d.

Here, WordNet (Fellbaum 1998) is employed as a semantic database. WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. As opposed to a normal dictionary, WordNet is a lexical ontology that attempts to organize lexical information hierarchically in terms of word meanings. Developed by hand over the past 17 years, it is a massive compilation of hierarchical relationships between each sense of every noun, verb, adjective and adverb in standard English. The 80 000 noun senses, for example, are organized into 11 hierarchies rooted by general concepts such as *entity*, *abstraction*, *human activity*, *event* and *group*. A word sense *C* may have any number of *subordinates*, which are specific types of *C*. For example, *object* is a subordinate of *entity*, and *natural-object*, *artefact* and *living-thing* are subordinates of *object*. Verbs have a similar set of hierarchies organized on the same principle of subordination, for example, *walking*

is one way to *travel*. Each word also has a free-form, textual definition similar to a dictionary entry. These definitions are stored as strings without any indexing or structured content. The hierarchical structure of WordNet allows the formulation of efficient search algorithms. Starting at a very general root node, as for example, *artefact*, the visual information found in the entire scene is used to find a specific subordinate concept to describe a particular region.

One shortcoming of WordNet with respect to this particular application is that the definitions and glosses generally do not contain any descriptions of the objects' appearance. A chair, for example, is defined to be *a seat for one person, with a support of the back*. In order to address this problem we have tagged the concepts in WordNet with attribute values indicating whether they are visible, capable of motion and usually located indoors or outdoors. Topic attributes also indicate relevance to specific topics. The attribute information that describes the type of motion is more detailed. We differentiate between *no motion*, *continuous motion* and *random motion*. In the case of continuous motion we also specify the dominant direction of the motion. So, for example, the dominate direction of climbing is upward or vertical. Currently these attributes are added by hand to the database. The hierarchical structure of WordNet enables very efficient attribution, as the subordinates inherit attribute values. For our experiments, *ca.* 195 000 attribute values are assigned.

(a) *Visual analysis*

To initiate and conduct the semantic search, visual observables were extracted from the video. The challenge here is to find a set of visual observables. These were then mapped into language terms that aid the semantic search of WordNet. Visual analysis can produce information at a large range of semantic levels. The sole fact that the scene contains a set of parallel lines by itself is not very helpful for the visual search. The set of parallel lines, how-

Figure 13. The visual information hierarchy induced by semantic levels.

ever, could indicate the presence of a man-made object. It is tempting to assume that image analysis can produce high-level semantic content, and map this information directly into WordNet. However, this is not feasible. For narrow, restricted problem domains, computer vision can indeed produce semantic content at the level of language. When the scope is a bit wider, as for example, in the broadcast video domain, the variety of objects and events present requires the development of tens of thousands of specialized object detectors.

Instead, our guiding principle was to use a small number of high-level object detectors, coupled with mid-level visual observables, to interpret the scene. We demonstrate later that by making extensive use of the semantic knowledge base and contextual information, a high-level scene description can be inferred from this combination of visual features. To formalize the semantic content of visual information, the extraction of visual descriptors was organized hierarchically (see figure 13). Edges, interest points, responses to filter banks, and segmented regions were considered to be *direct observables* and therefore form the basis of the visual information hierarchy (level 0). At level 1, information derived directly from image observables was computed and characterized. Level 2 consists of objects and events, which are recognized using data from the previous levels. Level 0 information is 2D, but carries little semantic information on its own. Level 1 information is more salient, but generally requires additional context to be meaningful. For the semantic search, level 2 information corresponds directly to semantic concepts in WordNet, while level 1 data are used to refine the search. We use three types of visual information across levels 1 and 2: people and their movements, other moving objects (e.g. vehicles) and stationary scene elements, such as man-made objects and structures, and natural objects, such as sky, water and vegetation. People are detected using a face detector courtesy of Carnegie Mellon University (Rowley *et al.* 1998), although we also plan to integrate human body detectors in future work. Our scene and motion descriptors are now described in more detail.

(i) Scene characterization

The goal of scene characterization is to semantically categorize every object and element in the scene. This pro-

visual observ- able	level	elemental	pruning	attribute
man-made	2	artefact, n, 1		
face	2	person, n, 1	person, human	
sky	2		sky, air	outdoors
motion	2	travel, v, 1	motion	
motion speed	1			fast, slow
motion direction	1			horizontal, up, down
motion type	1			smooth, erratic

Figure 14. Selected mappings from visual observables to semantic terms. Elemental terms have a part of speech (n, noun or v, verb), and a sense number in WordNet. Motion maps to the elemental term travel, and motion characteristics map to quantized attribute values.

vides a starting point for the semantic search to add further specialization or detail to the categorization, and provides scene context to constrain the search. This characterization is a very difficult problem, however, as semantic context may be required even at this stage. Our approach was to define a small set of categories, some of which were very broad, that could be distinguished primarily by 2D appearance information. The categories spanned virtually all objects of interest except for people, which were handled with specialized detectors. Currently, our categories are *man-made objects*, *vegetation*, *water/sea/ocean*, *rock* and *sky*. Although the man-made category is very large, it corresponds to the 'artefact' hierarchy in WordNet, which contains all man-made objects. The presence of certain categories in a scene, such as water and sky, also provided evidence of scene context such as *outdoors*, which was useful in the semantic search as described later. Algorithms specifically designed for overall scene classification (Szummer & Picard 1998; Vailaya *et al.* 1998) could be easily integrated into the system, as well as other categories such as animals.

The categorization was formulated as an integrated problem of perceptual grouping, texture recognition and segmentation. While filter banks are useful for distinguishing between similar textures, the salient features of these broad texture categories are often embedded in the relationships between large-scale features such as regions. In previous work we have examined this problem (Hoogs *et al.* 2003), and developed a region-based approach for texture characterization. The algorithm outperforms filter banks when the range of appearance within each texture is large compared to the training images, as is generally the case in broadcast video.

The approach has been extended to video, using a maximum-likelihood formulation to combine texture class estimates across subsequent video frames after inter-frame homographies are estimated. At this stage, the scene is assumed to be static, or to contain only slowly moving objects. Example results are shown in figures 15 and 18. Motion analysis is then applied to segment and characterize independently moving objects.

(ii) Motion characterization

Our system currently uses an extended version of the blob tracking approach presented in Comaniciu *et al.*

region 2, nouns:
water, crude oil: {oil tanker, oiler, tank ship, tanker}
sea, water: {mildam}, {Suez canal}, {abandoned ship, derelict}, {wreck},
{lifeboat}, {whaleboat}
water: {launch}, {frigate}, {bottom, freighter, merchant ship, merchant-man}

Figure 15. The original image (a) is segmented into regions (b) classified as ‘water’ (blue), ‘sky’ (white) and ‘man-made’ (light brown). The man-made region (region 2) is contained within the water region. The top 10 concepts returned by the man-made search are shown in rank order, grouped by matched constraint terms shown in bold. The words listed within each set of brackets are synonyms and represent a single concept.

clip 1	clip 2	clip 3	clip 4	clip 5	clip 6

water, sky, rock <i>none</i>	man-made, rock <i>none</i>	man-made, water, rock <i>none</i>	man-made, rock <i>none</i>	man-made, water, person manager, chairperson	man-made <i>none</i>
clip 7	clip 8	clip 9	clip 10	clip 11	clip 12

water, man-made, (person) lifeboat, launch; voy- ager, oceanographer	man-made, water launch, frigate, freighter	man-made <i>none</i>	man-made, water, sky helicopter	man-made, water, rock blacktop, tarmac	man-made, rock <i>none</i>

Figure 16. A complete story containing 12 clips was analysed. One keyframe for each of the clips is shown with the extracted scene characteristics (above the line) and any correct semantic annotations in the top 10 found by the search.

(2000). The motion trajectory is computed based on the motion of the centroid of the foreground object. In the context of gesture analysis, suitable state space representations (Bobick & Wilson 1997) have been used to recognize a limited vocabulary of hand gestures. Rather than extracting a mathematical description of the motion trajectory we are interested in its characteristic. Is it, for

example, a smooth horizontal motion or an erratic motion, which could indicate some kind of jumping or bouncing motion? Depending on how much we know about the camera and the scene, the motion trajectory allows us to estimate the speed of the motion, which can be a very important clue.
The discussion of recognizing motion patterns presented

Figure 17. Percentage of the top 10 concepts returned by the semantic search that are relevant to each clip.

in § 3 is of course highly relevant to this work. As opposed to the motion trajectory it is a local characteristic of the observed motion. We described techniques to recognize biological motion based on the spatio-temporal features of the image sequence. The structure and type of motion pattern also reveal whether a rigid or non-rigid object is being tracked. Furthermore, the periodicity of the motion can be computed. Polana & Nelson (1997) used a set of basic descriptors to recognize certain motion patterns. An approach to detect the periodicity of the motion is given in Liu & Picard (1998). Modal analysis (see figure 2) provides another possibility to analyse the observed motion. The presence of harmonic modes and their time constants β

indicates periodic motion. Currently the system uses only the dominant direction of the motion.

(b) Mapping from vision to language

To relate the disparate representations of visual observables and WordNet, we have defined the visual information hierarchy described in § 4a, and established the relationship of each level to WordNet concepts. This formalism provides a grounded framework to contain visual information, linguistic information and their respective uncertainties and ambiguities. Specifically, level 2 observables corresponding to objects of interest are mapped directly to general concepts in WordNet, and become *elemental terms* (see figure 14). This is possible because the semantic meaning of each level 2 observable is clearly defined, and can be mapped directly to a word sense. Such observables of interest include human faces and man-made objects, which are mapped to nouns. Moving objects, including moving people, are mapped to verbs. The remaining level 2 observables are used as contextual search constraints as described below.

Elemental terms are very general, and provide entry points for searching WordNet. To find more specific concepts present in the video, level 1 observables are mapped to *pruning terms*, or words that are used to constrain the search as described below. Unlike level 2 observables, however, it is often impossible to map each level 1 observable to a single word sense because of the well-known difficulty of consistently mapping low-level, quantitative data onto symbolic terms. This problem is partially addressed by mapping some level 1 observables to multiple words that are synonyms.

Certain visual observables are mapped to attributes described earlier. Level 1 observables in this category include

Figure 18. The original image (a) is segmented into regions (b) classified as 'rock' (red) and 'sky' (white), providing the attribute term 'outdoors'. The motion of the person is tracked and characterized as motion attribute terms 'upward' and 'slow'. Both a noun and a verb search are conducted using *person* and *human motion* as elemental terms. The most likely concepts returned by each search are shown in rank order, grouped by matched constraint terms shown in bold.

specific motion characteristics such as direction and speed. Some level 2 observables also map to attributes; for example, *sky* and *vegetation* map to the outdoors attribute. These level 2 observables do not become elemental terms, as we are not interested in categorizing them further in this work. Instead they provide scene context to constrain the search. Topics are mapped to both pruning terms and attribute terms; only a limited number of general topics are also attributes, as these are determined manually.

(c) *Semantic search*

As mentioned earlier, the objective of the semantic search is to find the most specific description of the scene elements supported by the evidence and topic context. Even with all observables in a common representation, there are still significant issues in searching WordNet. For example, the only difference between subordinates of the same node are their definitions, which are unstructured text. Natural language processing is required to extract information from definitions, and this process can be error prone. The length of the present article does not allow a detailed description of the search algorithm, but the main idea is to use the pruning information to find the most relevant concepts in the database. As a result of the search we obtain a ranked list of concepts. As will be seen later, these lists can of course contain concepts that have little or no relevance. The aim of the current work is to ensure relevant words are contained in the lists and have a sufficiently high ranking.

(d) *Video annotation results*

The purpose of our video annotation system is to derive semantic content descriptions. A video is currently segmented into clips (by hand in these experiments), and these clips are passed into the system. The scene characterization algorithm classifies these clips into the various categories. The face detector is applied in parallel, and supersedes segmentation results when humans are found. Next, the tracking algorithm, which is manually initialized in the current system, is applied to man-made regions and faces, to determine their motion. This visual information is translated into elemental, pruning and attribute terms as described in § 4b. Simultaneously, topic terms are extracted from the transcript (this step is currently manual). Finally, semantic search is performed to determine the annotations. Once the search terminates, the 10 highest ranked concepts are output so that we can report performance measures as described below.

The first example, shown in figure 15, illustrates how the system makes use of the scene context. Three regions are segmented and classified. Region 1 (blue) is classified as 'water', region 3 erroneously as 'sky', and region 2 (light brown) as 'man-made'. Region 2 is contained within region 1. The topic information is 'crude oil', which is mentioned repeatedly in the commentary; this becomes a pruning term, but not a topic attribute because it is too specific. The man-made region maps directly to the elemental term 'artefact' and initiates a search through the 14 000 subnodes of artefact. The pruning terms 'sea' and 'water' are added because the man-made region is contained in the water region. As demonstrated by the output noun list, using relatively simple visual observables and basic topic information, the correct object label of {oil

tanker, oiler, tank ship, tanker} is at the top of the ranked list. Without the 'crude oil' topic, the various types of boats are still detected as the most likely concepts, but with approximately equal likelihood.

Before discussing specific examples containing biological motion we present results of an initial study examining the system's ability to annotate a wide range of scenes. All clips in a contiguous news story about the *Exxon Valdez* oil spill were analysed. Typically, a 2–3 min story contains dozens of clips showing a wide range of subjects. Key frames from the clips of the story are shown in figure 16. Note the variety of scales, object types and scene content despite the common topic. No topic information was used in the search, however, as we were interested in the generic performance of the system.

Performance evaluation of this system has two levels. First, we assess the accuracy of the computed visual information. Then, we assess whether the semantic search found annotations that correctly improve upon the visual information. To measure the second stage, we manually examine the generated word lists to determine which words are appropriate descriptions of their corresponding scene elements. Currently we do not evaluate how specific the detected labels are. The system should be rewarded for finding more specific terms, and punished for finding overly specific terms. This will be addressed in future experiments.

The percentage of correct labels in the top 10 words is shown in figure 17. The extracted visual information is listed below each image above the line; the correct annotations found by the search are listed below the line. On approximately half of the clips, the semantic search found relevant annotations among the top 10 words. On the other half, one of two conditions applies. In some cases, such as the first clip, no objects corresponding to elemental terms were detected. Hence no search was performed, and the system output only visual analysis. In other cases, elemental terms were detected, but the additional visual evidence was insufficient to specialize the objects further. This occurred on clips 2, 3, 4, 6, 9 and 12. Another problem was of course errors caused by the segmentation. The detection of water in clips 10 and 11 was erroneous. In these particular examples, however, the system still produced the appropriate labels.

Finally, we discuss two examples that contain biological motion. In the example shown in figure 18 a man climbs a rock. A visual tracker is used to track the motion of the person (initialized manually in this example). The scene is segmented and classified into two regions, 'rock' and 'sky'. The person is misclassified as 'rock', but this error can be corrected when motion is present or the person is detected. The semantic search is conducted starting at the noun 'person', using the attribute term 'outdoors' derived from 'sky' and the pruning term 'rock'. This yields a list of nouns that contains the correct object label {rock climber}. The false positive 'punk rocker' is due to the ambiguity of the meaning of the word 'rock' in WordNet definitions (which are untagged strings). The verb search is conducted from the verb concept 'human motion', and constrained by the extracted motion attributes 'up' and 'slow'. This yields a verb list including {ascend, come up, rise, uprise}, {scale} and {escalate}. The verb {climb} was also found, but was ranked just below the top 10 verbs.

This illustrates the potential benefits of considering noun and verb searches jointly, which is a topic for future research. But in principle this example shows that the use of a semantic database can be useful, and frees us from learning very specific scenarios.

In § 3 we detected the running motion (see figure 9) using the spatio-temporal pattern of the image sequence. Based on the motion of the centroid of the foreground blob we know that it is a fast motion in the horizontal direction. The semantic verb search using the pruning information *horizontal*, *articulated* and *fast* returns the words: march, pace, outride, romp, run, outrun, sprint and rush. Most of these words are an adequate description of the observed motion.

Our experiments show that with the existing set of visual analysis algorithms—face detector, scene characterization, motion tracking and semantic search—we can produce meaningful, non-trivial annotations on a wide variety of scenes. The results shown here are representative of our other experiments, and give an indication of the potential of the approach. Obviously, more experimentation needs to be performed, particularly to examine the trade-offs between topic context, the amount and levels of visual information and the specificity of the search. Currently, the system is limited by the visual algorithms, notably the lack of a person detector, motion track initialization and the segmentation algorithm. As stated earlier, information of the motion pattern is currently not used.

Although WordNet was chosen because of its hierarchical structure we also need to consider other semantic databases. FrameNet (Baker *et al.* 1998), invented by C. Fillmore, for example, is based on the idea of semantic frames. Having more information of the possible or most probable noun-verb combinations will improve the analysis of events. Another idea is also to make extensive use of the audio information. Background noises, as for example, engine noise, could be processed in a similar fashion to the way that the visual information is used now. A number of researches have also collected co-occurrence information of words. This information would be useful to improve our current ranking. It would, for example, help to disambiguate the sense of rock in figure 18.

5. CONCLUSION

The aim of the work presented in this article is to explore mathematically rigorous approaches to recognize and classify biological motion. Guided by the principle that perception and classification should not be dealt with as two separate processes, the contour tracking framework was used to simultaneously track and classify the motion. We demonstrated how effectively ARPs can mimic certain biological motions (see figure 3). These learnt motion models could then be used as observation densities in an HMM. Other than in Bregler (1997), the discrete state variable of the HMM is not used for modelling the motion and can be used for classification. An automatic segmentation of a sequence containing more than one type of motion is demonstrated in figure 4. It is demonstrated that the Markov chain can be applied more effectively by using partial importance sampling.

Although the fusion of classification and perception itself is successful we concluded that the contour tracking

framework is not a very robust method for tracking human beings. Apart from the problem of initialization, finely tuned motion models are needed for the classification of the type of motion. It would be advantageous to recognize a particular motion pattern directly from the spatio-temporal features in the image sequence, and not use any form of intermediate representation such as a spline contour or an articulated model. The method of extracting skew vectors from a set of epipolar slices turns out to be a very effective feature extraction method. The experiments demonstrate that the automatic segmentation of sequences containing more than one type of motion was successful. Now we are able to analyse more complex sequences, as for example, the basketball sequence shown in figure 12. One shortcoming of the new method is also that the skew vectors are not a viewpoint-invariant motion descriptor. This and a more systematic approach to the low-level feature extraction itself are the main topics for future research.

Finally the approach presented in § 4 illustrates that it is possible to generate a higher-level description of the scene automatically. The examples in figure 18 show that a small number of low-level visual descriptors is sufficient to label the type of motion. This approach does not rely on object-specific or event-specific learning and has the potential to scale up to a large problem domain. The objective of our current research is to extend the system's capability to recognize certain biological motions using the context information of the scene.

The work presented in §§ 2 and 3 was carried out by J.R. and A.B. at the University of Oxford. J.R. was supported by a Marie-Curie fellowship awarded by the European Union. The work presented in § 4 was carried out by J.R., A.H. and G.S. at GE Global Research. This work was supported in full by the Advanced Research and Development Activity (ARDA). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US Government. Furthermore we thank Steve Roberts (University of Oxford), Joseph Mundy (Brown University), George A. Miller and Christiane D. Fellbaum (Princeton University) for fruitful discussions and guidance and John Schmiederer for implementation and experimental support.

REFERENCES

- Baker, C. F., Fillmore, C. J. & Lowe, J. B. 1998 The Berkeley FrameNet project. In *Proc. 36th A. Mtng of the Assoc. for Comput. Linguistics and 17th Int. Conf. on Comput. Linguistics* (ed. C. Boitet & P. Whitelock), pp. 86–90. San Francisco, CA: Morgan Kaufmann Publishers.
- Baumberg, A. & Hogg, D. 1994 Learning flexible models from image sequences. In *Proc. 3rd European Conf. Computer Vision, Stockholm, Sweden*, pp. 299–308. Stockholm, Sweden: Springer.
- Black, M. J. & Jepson, A. D. 1998 A probabilistic framework for matching temporal trajectories: condensation-based recognition of gestures and expressions. In *Proc. 5th European Conf. Computer Vision, Freiburg, Germany*, pp. 909–924.
- Blake, A. & Isard, M. 1998 *Active contours*. Berlin: Springer.
- Bloom, P. 2000 *How children learn the meaning of words*. Cambridge, MA: MIT Press.
- Bobick, A. F. & Wilson, A. D. 1997 A state-based approach to the representation and recognition of gesture. *IEEE Trans. Pattern Analysis Machine Intell.* **19**, 1325–1337.

- Borgefors, G. 1988 Hierarchical chamfer matching: a parametric edge detection algorithm. *IEEE Trans. Pattern Analysis Machine Intell.* **10**, 849–865.
- Bregler, C. 1997 Learning and recognizing human dynamics in video sequences. In *Proc. 11th IEEE Computer Vision and Pattern Recognition, San Taun, PR*, pp. 568–574.
- Bülthoff, I. & Bülthoff, H. 2003 *Analytic and holistic processes in the perception of faces, objects, and scenes, chapter image-based recognition of biological motion, scenes and objects*. Norwood, NJ: Ablex Publishing Corporation.
- Cipolla, R. & Yamamoto, M. 1990 Stereoscopic tracking of bodies in motion. *Image Vision Computing* **8**, 85–90.
- Comaniciu, D., Ramesh, V. & Meer, P. 2000 Real-time tracking of non-rigid objects using mean shift. In *IEEE Computer Vision and Pattern Recognition, Hilton Head, SC*, vol. 2, pp. 142–149.
- Deutscher, J., Blake, A. & Reid, I. 2000 Articulated body motion capture by annealed particle filtering. In *IEEE Computer Vision and Pattern Recognition, Hilton Head, SC*, pp. 126–133.
- Fellbaum, C. (ed.) 1998 *WordNet*. Cambridge, MA: MIT Press.
- Gavrila, D. 2000 Pedestrian detection from a moving vehicle. In *Proc. 6th European Conf Computer Vision, Dublin, Ireland*, pp. 37–49.
- Gavrila, D. & Philomin, V. 1999 Real-time object detection for smart vehicles. In *Proc. 7th Int. Conf. on Computer Vision, Corfu, Greece*, pp. 87–93.
- Gelb, A. (ed.) 1979 *Applied optimal estimation*. Cambridge, MA: MIT Press.
- Gordon, N., Salmond, D. & Smith, A. 1993 Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proc.* **140**, 107–113.
- Grenander, U. 1976–1981 *Lectures in pattern theory I, II and III*. Berlin: Springer.
- Hammersley, J. & Handscomb, D. 1964 *Monte Carlo methods*. London: Methuen.
- Hoogs, A., Collings, R., Kaucic, R. & Mundy, J. 2003 A common set of perceptual observables for grouping, figure-ground discrimination and texture classification. *IEEE Trans. Pattern Analysis Machine Intell.* **25**(4).
- Isard, M. & Blake, A. 1996 Contour tracking by stochastic propagation of conditional density. In *Proc. 4th European Conf. Computer Vision, Cambridge, UK*, vol. 1, pp. 343–356.
- Johansson, G. 1976 Visual motion perception. *Sci. Am.* **232**, 76–89.
- Karatzas, I. & Shreve, S. E. 1991 *Brownian motion and stochastic calculus*. Berlin: Springer.
- Kass, M., Witkin, A. & Terzopoulos, D. 1987 Snakes: active contour models. In *Proc. 1st Int. Conf. on Computer Vision, London, UK*, pp. 259–268.
- Kitagawa, G. 1996 Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graphical Statist.* **5**, 1–25.
- Liu, F. & Picard, R. W. 1998 Finding periodicity in space and time. In *Proc. 6th Int. Conf. on Computer Vision, Bombay, India*, pp. 376–383.
- Lütkepohl, H. 1993 *Introduction to multiple time series analysis*, 2nd edn. Berlin: Springer.
- Niyogi, S. A. & Adelson, E. H. 1994a Analyzing and recognizing walking figures in xyt. In *Proc 9th IEEE Computer Vision and Pattern Recognition, Seattle, WA*, pp. 469–474.
- Niyogi, S. A. & Adelson, E. H. 1994b Analyzing gait with spatiotemporal surfaces. In *Workshop on Non-Rigid Motion and Articulated Objects, Austin, TX*.
- North, B., Blake, A., Isard, M. & Rittscher, J. 2000 Learning and classification of complex dynamics. *IEEE Trans. Pattern Analysis Machine Intell.* **22**, 1016–1034.
- Pearl, J. 1988 *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann Publishers.
- Polana, R. & Nelson, R. 1997 Detection and recognition of periodic, non-rigid motion. *Int. J. Computer Vision* **23**, 261–282.
- Rabiner, L. & Bing-Hwang, J. 1993 *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.
- Rittscher, J. & Blake, A. 1999 Classification of human body motion. In *Proc. 7th Int. Conf on Computer Vision, Corfu, Greece*, vol. 2, pp. 634–639.
- Rowley, H. A., Baluja, S. & Kanade, T. 1998 Neural network-based face detection. *IEEE Trans. Pattern Analysis Machine Intell.* **20**, 23–38.
- Schaal, S., Ijspeert, A. & Billard, A. 2003 Computational approaches to motor learning by imitation. *Phil. Trans. R. Soc. Lond. B* **358**, 537–547. (DOI 10.1098/rstb.2002.1258.)
- Scheffe, H. 1959 *The analysis of variance*. New York: Wiley.
- Schroeder, W. J., Zarge, J. A. & Lorensen, W. E. 1992 Decimation of triangle meshes. *Computer Graphics* **26**, 65–70.
- Sidenbladh, H., Black, M. J. & Fleet, D. J. 2000 Stochastic tracking of 3D human figures using 2D image motion. In *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, vol. 2, pp. 702–718.
- Szummer, M. & Picard, R. W. 1998 Indoor-outdoor image classification. In *IEEE International Workshop on Content Based Access of Image and Video Databases, in conjunction with ICCV'98*, pp. 42–51.
- Toyama, K. & Blake, A. 2001 Probabilistic tracking in metric space. In *Proc. 8th Int. Conf. on Computer Vision, Vancouver, BC*, vol. 1, pp. 50–57.
- Vailaya, A., Jain, A. & Zhang, H. 1998 On image classification: city vs. landscape. In *IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*, pp. 3–8.
- Zelnik-Manor, L. & Irani, M. 2001 Event-based analysis of video. In *IEEE Computer Vision and Pattern Recognition, Hawaii*.

GLOSSARY

ARP: autoregressive process
HMM: hidden Markov model