Chapter 2
Behaviour in Context

Interpreting behaviour from object action and activity is inherently subject to the context of a visual environment within which action and activity take place. Context embodies not only the spatial and temporal setting, but also the intended functionality of object action and activity (Bar 2004; Bar and Aminoff 2003; Bar and Ullman 1993; Biederman et al. 1982; Palmer 1975; Schwartz et al. 2007). Humans employ visual context extensively for both effective object recognition and behaviour understanding. For instance, one recognises, often by inference, whether a hand-held object is a mobile phone or calculator by its relative position to other body parts such as closeness to the ears, even if they are visually similar and partially occluded by the hand. Similarly for behaviour recognition, the arrival of a bus in busy traffic is more likely to be inferred by looking at the passengers' behaviour at a bus stop. Computer vision research on visual analysis of behaviour embraces a wide range of studies on developing computational models and systems for interpreting behaviour in different context. Here we highlight some well established topics and emerging trends.

2.1 Facial Expression

Behaviour exhibited from a human face is predominantly in the form of facial expression. The ability to recognise the affective state of a person is indispensable and important for successful interpersonal social interaction. Affective arousal modulates all non-verbal communication cues such as facial expressions, body postures and movements. Facial expression is perhaps the most natural and efficient means for humans to communicate their emotions and intentions, as communication is primarily carried out face to face.

Automatic facial expression recognition has attracted much attention from behavioural scientists since the work of Darwin (1872). Suwa et al. (1978) made the first attempt to automatically analyse facial expressions from images. Much progress has been made in the last 30 years towards computer-based analysis and interpretation of facial expression (Bartlett et al. 2005; Chang et al. 2004; Cohen et al. 2003;

S. Gong, T. Xiang, *Visual Analysis of Behaviour*,
DOI 10.1007/978-0-85729-670-2_2, © Springer-Verlag London Limited 2011

Donato et al. 1999; Dornaika and Davoine 2005; Essa and Pentland 1997; Fasel and Luettin 2003; Hoey and Little 2004; Jia and Gong 2006, 2008; Kaliouby and Robinson 2004; Lee and Elgammal 2005; Lyons et al. 1999; Pantic and Patras 2006; Pantic and Rothkrantz 2000, 2004; Shan et al. 2009; Yacoob and Davis 1996; Yeasin et al. 2004; Yin et al. 2004; Zalewski and Gong 2005; Zhang and Ji 2005). A recent trend in human–computer interaction design also considers the desire for affective computing in order to bring about more human-like non-verbal communication skills to computer-based systems (Pantic and Rothkrantz 2003).

Facial expressions can be described at different levels. A widely used description is the Facial Action Coding System (FACS) (Ekman and Friesen 1978), which is a human-observer-based protocol developed to represent subtle changes in facial expressions. With FACS, facial expressions are decomposed into one or more action units. The aim to develop computer-based automatic AU recognition has attracted much attention in the last decade (Donato et al. 1999; Tian et al. 2001; Valstar and Pantic 2006; Zhang and Ji 2005). Psychophysical studies also indicate that basic emotions have corresponding universal facial expressions across all cultures (Ekman and Friesen 1976). This is reflected by most facial expression recognition models attempting to recognise a set of prototypic emotional expressions, including 'disgust', 'fear', 'joy', 'surprise', 'sadness' and 'anger' (Bartlett et al. 2005; Cohen et al. 2003; Lyons et al. 1999; Yeasin et al. 2004; Yin et al. 2004). Example images of different facial expressions are shown in Fig. 2.1.

Although a facial expression is a dynamic process with important information captured in motion (Bassili 1979), imagery features extracted from static face images are often used to represent a facial expression. These features include both geometric features (Valstar and Pantic 2006; Valstar et al. 2005) and appearance features (Bartlett et al. 2005; Donato et al. 1999; Lyons et al. 1999; Tian 2004; Zhang et al. 1998). The rigidity and structural composition of facial action components make facial expression a rather constrained type of behaviour, a raison d'être for extracting information from static images to discriminate different types of facial expression. Efforts have also been made to model facial expression as a dynamical process utilising explicitly spatio-temporal or image sequence information (Cohen et al. 2003; Shan et al. 2005; Zalewski and Gong 2005; Zhang and Ji 2005). These studies aim to explore a richer description of expression dynamics so to better interpret facial emotion and expression change.

2.2 Body Gesture

The ability to interpret human gestures constitutes an essential part of our perception. It reveals the intention, emotional state or even the identity of the people surrounding us and mediates our communication (Pavlovic et al. 1997). Such human activities can be characterised by hand motion patterns and modelled as trajectories in a spatio-temporal space. Figure 2.2 shows some examples of communicative gestures. Gesture recognition usually considers the problem of measuring the similarities between hand motion pattern characteristics, such as their trajectories, and

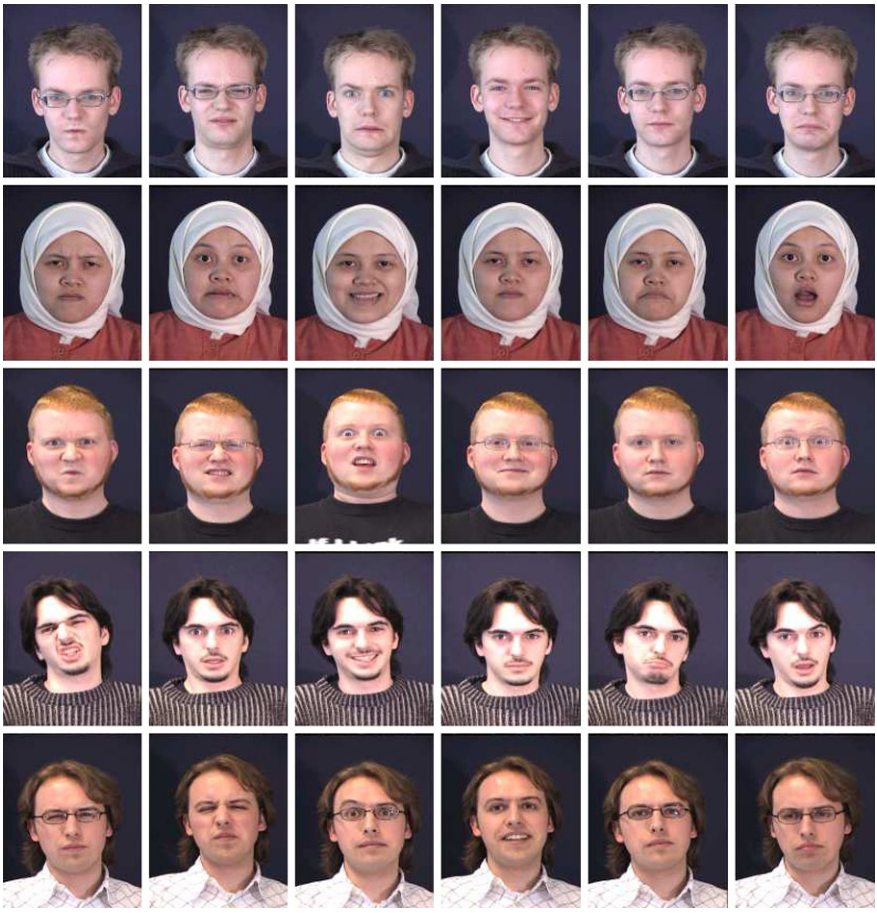


Fig. 2.1 Examples of facial expression images from the MMI database (Pantic et al. 2005)

discriminating different types of such patterns based on an assumption that each type of gesture exhibits a unique or closely similar trajectory with small variations. Based on this simple concept, well-defined gestures can be recognised by computer-based models through matching trajectory templates (Davis and Bobick 1997; McKenna and Gong 1998; Wren et al. 1997). In a well-constrained environment with little background clutter where only a single or a dominant person is assumed, for example, in computer games and close-range human–computer interaction, a holistic gesture representation can be computed without explicit body part tracking. For example, motion moment measurements projected onto the image plane (camera view-plane) axes can be used for a hand-driven interface (Freeman et al. 1996). However, most natural gestures are not well-defined in the sense that there is no clear starting or ending point for a gesture, and multiple gestures form a single continuous pattern with possible abrupt interruptions. Different gestures can also display visually similar motion patterns. To address these problems, gestures

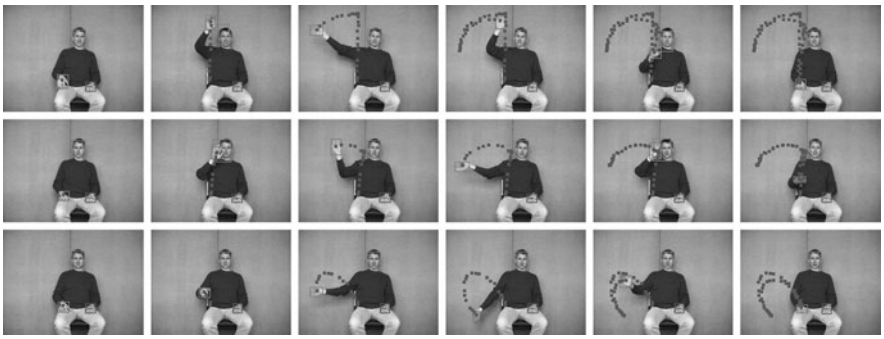


Fig. 2.2 Examples of communicative gestures: 'waving high' (*top row*), 'waving low' (*middle row*) and 'please sit down' (*bottom row*)



Fig. 2.3 Examples of affective body gestures from the FABO database (Gunes and Piccardi 2006): 'fear' (*1st row*), 'joy' (*2nd row*), 'uncertainty' (*3rd row*), and 'surprise' (*4th row*)

can be modelled as stochastic processes under which salient phases are modelled as state transitions. Prior knowledge is built into the models statistically by estimating both state distributions and visual observation covariance at each state, and can be learned from previous observations of similar and different gesture motion patterns, known as learning from training data (Black and Jepson 1998; Bobick and Wilson 1995; Gong et al. 1999; Psarrou et al. 1995, 2002; Schlenzig et al. 1994; Starner and Pentland 1995).

Both human face and body contribute collectively in conveying the emotional state of a person (see examples in Fig. 2.3). The perception of facial expression is strongly influenced by body motion. This is probably due to a high variability in body movement when visually observed, as body language is defined collectively by both body movements including posture (static), gesture (dynamic), and facial movement including expression and gaze patterns. Studies in psychology suggest

that the combined visual channels of facial expression and body gesture are more informative than each alone, and their integration is a mandatory process occurring early in the human cognitive process (Ambady and Rosenthal 1992; Meeren et al. 2005). Coulson (2004) shows experiments on attributing six universal emotions to static body postures using computer-generated mannequin figures. These experiments suggest that recognition from static body posture alone is comparable to recognition from voice, and some postures can be recognised as accurately as facial expressions.

Early studies on vision-based gesture recognition were primarily carried out on non-affective gestures, such as sign languages (Wu and Huang 2001). An affective gesture recognition system was introduced by Ravindra De Silva et al. (2006) to recognise children's emotion with a degree of intensity in the context of a game. More recent studies on affective modelling consider both facial expression and body gesture (Balomenos et al. 2005; Gunes and Piccardi 2007; Shan et al. 2007). Kapoor and Picard (2005) describe a multi-sensor affect recognition system for classifying the affective state of interest in children who are solving puzzles. This system combines the extracted sensory information from face videos (expression), a sensor chair (posture), and the state of the puzzle (context). Burgoon et al. (2005b) consider the problem of identifying emotional states from bodily cues for human behaviour understanding. Statistical techniques are often exploited to estimate a set of posture features in discriminating between emotions (Ravindra De Silva and Bianchi-Berthouze 2004). Overall, how to effectively model jointly facial and body behaviour remains an open question.

2.3 Human Action

Automatic human action recognition has become a very active research area in computer vision in recent years with applications in human–computer interaction, video indexing, visual surveillance, sport event analysis and computer games. Automatic action recognition is challenging because actions, similar to human gestures, can be performed by people of different size, appearance, and from different viewpoints. The problem is compounded by occlusion, illumination change, shadow, and possible camera movement. In general, action recognition considers a less constrained visual environment compared to that of gesture recognition. Some examples of actions can be seen in Fig. 2.4.

Action recognition requires a suitable representation. An intuitive approach to representing action can be based on either tracking a spatio-temporal shape of a person (Ali and Aggarwal 2001; Ramanan and Forsyth 2003; Rao and Shah 2001; Sheikh et al. 2005), or constructing a shape template and analysing its spatio-temporal characteristics (Gorelick et al. 2007; Ke et al. 2005; Yilmaz and Shah 2005; Zhang and Gong 2010a, 2010b). This spatio-temporal shape is often computed by extracting the silhouette of a person in motion, which may not be possible in a cluttered visual environment. To overcome this problem, less descriptive and shape-free representations can be considered. They include space-time interest points (Laptev

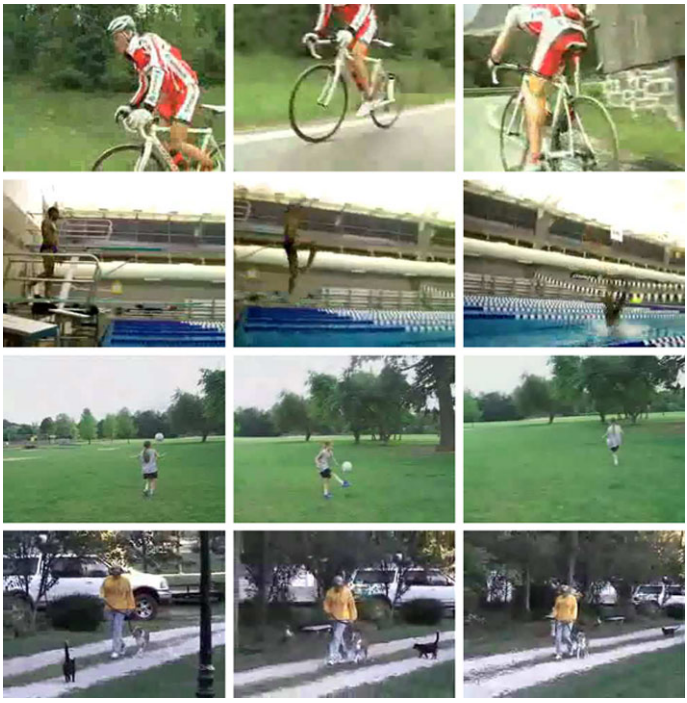


Fig. 2.4 Example actions in unconstrained environments. From *top* to *bottom*: 'cycling', 'diving', 'soccer juggling', and 'walking with a dog'

2005; Laptev and Lindeberg 2003; Schüldt et al. 2004), optical flow (Efros et al. 2003), motion template (Bobick and Davis 2001), space-time volumes (Yilmaz and Shah 2005), and shape context from still images (Belongie et al. 2002; Wang et al. 2006a).

Compared to other methods, a histogram representation of space-time interest point is particularly robust against noise and partially missing object features. However, a weakness of this representation is that the overall object spatio-temporal structural information is lost due to its histogram nature. A number of recent studies have tried to exploit the use of spatial or temporal structural information about interest points through a correlogram (Liu and Shah 2008), spatial location registration (Gilbert et al. 2008), motion context (Zhang et al. 2008), and spatio-temporal selection (Bregonzio et al. 2009).

Most studies on action recognition are concerned with classifying actions into different types, making an assumption that isolated visual sequences of individual actions are readily detectable and segmented. However, realistic actions are often observed continuously in the presence of other object actions. In such an environment, action detection and segmentation become difficult. Current action detection models are rather simplistic, largely based on template matching, with an action template constructed from previously seen video clips. Significant challenges remain. For instance, Ke et al. (2007) consider a semi-supervised part-based template



(a) (b) (c)

Fig. 2.5 (**a**) Facial expression, (**b**) head pose, and (**c**) gait may be used as behavioural cues to reveal human intent

for action detection. Whilst only a single training sample per action class is required, intensive user interaction including manual segmentation of body parts in each image frame is required for developing a part based shape-flow model. Alternatively, Yang et al. (2009) exploit an auxiliary annotated action dataset to compensate for the lack of detection training data. To cope with large intra-class action variations in detection, weakly supervised model learning strategies have also been studied (Hu et al. 2009; Siva and Xiang 2010; Yuan et al. 2009).

2.4 Human Intent

It could be considered that the emotional state of a person can provide non-verbal visual indicators for gauging one's intent. However, interpretation of human intent by non-verbal cues alone is challenging (Burgoon et al. 2005a; Vrij 2000). Situational body language can be visually difficult to interpret, especially if there is a deliberate attempt to deceive. This is in contrast to communicating emotional states using heightened facial expression and affective gesture. Nevertheless, there is evidence to suggest that non-verbal cues can be correlated with intent, and in particular for facilitating the detection of deceptive intent (Vrij and Mann 2005). From a visual analysis perspective, one may consider detecting intent by modelling the correlations among a range of non-verbal behavioural cues captured from facial expression, eye movement, head pose changing patterns, gesture, and gait (Fig. 2.5).

Visually gauging one's intent from a far-field distance is more difficult than in a close-range environment due to a lack of visual details. In an unconstrained open environment, analysing facial expression and eye movement patterns is unreliable due to low resolution and occlusion. In contrast, body orientation and head pose profiles over space and time can provide more reliable measurements for correlating intent. For instance, on a train platform a person facing the train with his head turning towards another direction may indicate he is checking the information board for ensuring the correct train to board (see Fig. 2.6). Two people slightly tilting their heads towards each other may suggest they are chatting. Furthermore, body orientation and head pose of a person often differ, although they are intrinsically correlated

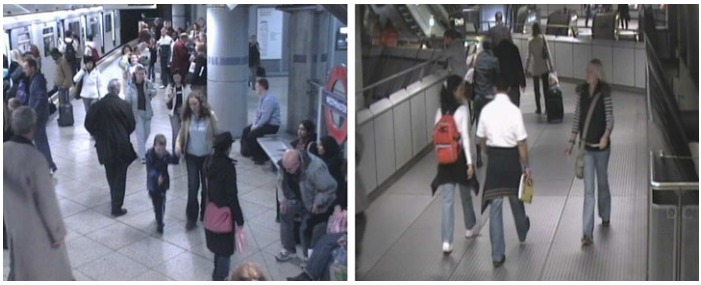


Fig. 2.6 Visual inference of intent in an unconstrained environment is challenging

due to physiological and behavioural constraints. Both can be visually estimated and analysed collaboratively for better accuracy and robustness.

There is a large body of research on automated human detection and estimation of body and head pose. However, most existing techniques concentrate on a well-controlled environment with high-resolution imagery data (Aghajanian et al. 2009; Andriluka et al. 2009; Gong et al. 1996; Jiang 2009; Lee and Cohen 2004; Ng and Gong 2002; Ozuysal et al. 2009; Sherrah and Gong 2001; Sherrah et al. 2001; Zhang et al. 2007). A few recent studies begin to address the problem of estimating head pose given low-resolution images (Benfold and Reid 2008, 2009; Robertson and Reid 2006) and in crowded scenes (Orozco et al. 2009). For estimating articulated body orientation, one can also exploit the mapping of kinematically constrained object parts to a pictorial structure of body parts appearance (Felzenszwalb and Huttenlocher 2005). Furthermore, joint head pose and body orientation can be estimated to yield a more holistic and reliable measurement towards modelling body language (Gong et al. 2010). In addition to head pose, body posture and orientation, human gait can be considered for intent interpretation. Gait analysis typically involves the characterisation of periodic motion and spatio-temporal patterns of human silhouettes for identity recognition (Bashir et al. 2010; Han and Bhanu 2006; Nixon et al. 2005). It could be argued that analysing correlations among a person's head pose, body orientation, posture and gait, in relation to camera positions and reactions to external events, all play a role for the interpretation of intent of a person in an unconstrained environment using non-verbal cues.

2.5 Group Activity

When a sequence of different actions executed by the same object continue over time, such patterns of behaviour are considered to be activities. The activities of a person in a kitchen could be making a sandwich or washing up after lunch. Each activity consists of a series of ordered actions. For instance, making a sandwich could be composed of opening the fridge to take out bread and fillings, making the sandwich, and wrapping up. When activities involve a group of objects interacting, understanding activities is more concerned with reasoning about inter-action and

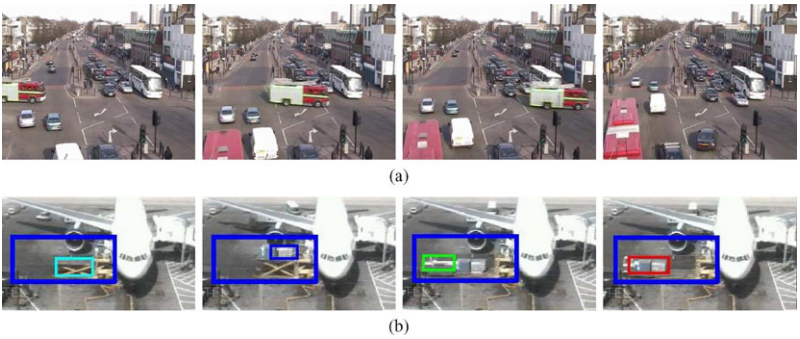


Fig. 2.7 Examples of group activities. (**a**) Different types of vehicles and pedestrians negotiate a traffic junction. (**b**) A cargo loading activity for a docked aircraft where a truck brought cargo container boxes onto a cargo lift, which then transferred the boxes into the aircraft. The cargo loading area and objects involved in the activities are highlighted by different bounding boxes

inter-person relationships including temporal order and duration among constituent actions of the individuals involved. Typical examples of group activities include passengers arriving at an airport, going through immigration desks and luggage collection; and vehicles and pedestrians negotiating a traffic junction regulated by traffic lights. Some examples of group activities can be seen in Fig. 2.7.

A fundamental element in understanding group activities is context. Visual context is the environment, background, and settings within which objects and associated behaviours are observed visually. Humans employ visual context extensively for both object recognition in a static setting and behaviour recognition in a dynamic environment. Extensive cognitive, physiological and psychophysical studies suggest that visual context plays a critical role in human visual perception (Bar and Aminoff 2003; Bar and Ullman 1993; Biederman et al. 1982; Palmer 1975). For visual analysis of behaviour, the most relevant visual context is behavioural context, as the interpretation of object behaviour is context dependent (Li et al. 2008b). In particular, a meaningful interpretation of object behaviour depends largely on knowledge of spatial and temporal context defining where and when it occurs, and correlational context specifying the expectation inferred from the correlated behaviours of other objects co-existing in the same scene. Modelling group activity can facilitate learning visual context, which in turn facilitates context-aware abnormal behaviour detection. For example, Li et al. (2008a) study the problem of detecting unusual behaviours that are visually similar to the norm when observed in isolation, but contextually incoherent when both spatial and correlational context are considered. There is a growing interest in modelling visual context for interpretation (Fernández et al. 2010; Kim and Grauman 2009; Kuettel et al. 2010; Loy et al. 2011; Streib and Davis 2010; Van Gool et al. 2009; Zaharescu and Wildes 2010).

Object tracking in a public space is a challenging problem, as object visual features are neither consistent nor unique (Gong and Buxton 1993). Understanding group activity can provide a form of dynamical context for better interpretation of individual object's whereabouts. For instance, Zheng et al. (2009) embed contextual



(a) Person re-identification in isolation is challenging due to visual ambiguities.

(b) Associating groups of people can reduce ambiguities in matching.

Fig. 2.8 Visual context extracted from surrounding people could be used to resolve ambiguities in person re-identification tasks

visual knowledge extracted from surrounding people into supporting descriptions for matching people across disjoint and distributed multiple views, known as person re-identification (see examples in Fig. 2.8). In another study, Ali and Shah (2008) exploit scene structure and behaviour of the crowd to assist appearance-based tracking in highly crowded scenes. More generally, tracking-by-detection models exploit either pedestrian or body part detection as categorical contextual information for more reliable tracking (Breitenstein 2009; Okuma et al. 2004; Wu and Nevatia 2007). There are also studies that exploit contextual information around a target object both spatially and temporally to facilitate more robust long-term tracking (Yang et al. 2008).

2.6 Crowd Behaviour

When the size of a group of objects increases to an extent that individual body movement cannot be distinguished, group activity becomes crowd behaviour. For instance, on a crowded train platform, people often move together in a synchronised pattern for departing or boarding a train. Visual analysis of behaviour of a single person becomes less reliable, and less relevant, whilst understanding the crowd behaviour as a whole is of more interest.

Crowd behaviour is characterised by severe occlusions due to excessive number of objects in the scene and a lack of visual details.[1] In particular, in a typical public scene as shown in Fig. 2.9, the sheer number of objects with complex activities cause severe and continuous inter-object occlusions, leading to temporal discontinuity in object movement trajectories. Tracking can be further compromised by low frame rates in image capturing, where large spatial displacement of objects occur between consecutive image frames.

[1] Many surveillance systems record videos at less than 5 frames per second (FPS), or compromise image resolution in order to optimise data bandwidth and storage space (Cohen et al. 2006; Kruegle 2006).



Fig. 2.9 Object tracking in crowd is challenging due to significant visual ambiguities. Severe inter-object occlusion and lack of visual details at times are key factors that render object tracking infeasible

Many recent studies approach the problem of modelling crowd activity using alternative representations instead of object trajectories in order to circumvent unreliable tracking. Among them, pixel-based representation that extracts pixel-level features such as colour, texture and gradient, emerges as an attractive solution since it avoids explicit object segmentation and tracking, especially applicable for representing activity in crowded scenes.

Crowd analysis can benefit from utilising region based scene decomposition, whereby a wide-area scene is segmented into local regions, followed by a global activity modelling to discover and quantify the space-time dependencies between local activities. Without explicit object segmentation, this holistic approach is applicable to dense occluded crowds in contrast to methods relying on object-based decomposition. Different techniques have been proposed following this idea. The most common strategy is to divide a video into local fixed-size spatio-temporal volumes, within which local activity patterns are extracted (Hospedales et al. 2009; Kim and Grauman 2009; Kratz and Nishino 2009; Mehran et al. 2009; Wang et al. 2009) (Fig. 2.10(a)). There are also attempts for exploiting more rigid, top-down structural knowledge in decomposing a scene into exit and entry zones, paths, and junctions (Breitenstein et al. 2008; Makris and Ellis 2005; Wang et al. 2006b) (Fig. 2.10(b)) and semantic regions (Li et al. 2008b; Loy et al. 2010) (Fig. 2.10(c)).

2.7 Distributed Behaviour

Activities can cover large spaces over different time zones. For example, a person moves from his home to work across a city; or a passenger checks in an airline desk, passes through a security check and boards a plane. To understand and interpret such activities, a single viewpoint captured from one camera is inadequate. Behaviours



(a) (b) (c)

Fig. 2.10 Examples of activity representation by scene decomposition techniques: (**a**) local spatio-temporal volumes (Hospedales et al. 2009), (**b**) movement paths (*white areas*) and junctions (*black areas*) (Makris and Ellis 2005), and (**c**) semantic regions, each of which encapsulates similar activities regardless of object class and differs from those observed in other regions (Li et al. 2008b)

need be understood in a coherent global context observed from distributed multiple cameras. In comparison to a single camera system, the field of view of a multiple camera system is not limited by physical layout of a scene. It offers the potential for providing a complete record of an object's activity between spaces over time, allowing a global interpretation of the object's behaviour. The main characteristic of multi-camera behaviour correlation, in contrast to behavioural interpretation in a single camera view, is that there can be uncertain spatial and temporal gaps between camera views leading to missing information for a coherent interpretation of behaviours across all camera views.

In most distributed multiple camera systems, it is the norm to have disjoint fixed-position cameras with non-overlapping field of views. This is designed to maximise spatial coverage in a wide-area scene whilst minimising the deployment cost. Figure 2.11 shows an illustration of different types of overlap between multi-camera field of views. Apart from static cameras, there exists a combination of heterogeneous sensors such as pan-tilt-zoom (PTZ) cameras, mobile wearable cameras, and non-visual positioning sensors to jointly monitor and interpret activities in a wide-area scene of complex structures.

Significant efforts have been made in the last few years by the computer vision community towards developing multi-camera models, focusing on learning behaviour correlations across camera views. The problem of learning inter-camera relationships manifests itself most frequently as either the object association problem (Javed et al. 2003, 2005; Stauffer and Tieu 2003), or the topology inference problem (Makris et al. 2004; Tieu et al. 2005). The former emphasises more on associating cross-camera trajectory correspondence, whilst topology inference aims to infer a scene model for the camera topology of a camera network.

A more general problem for object association across camera views is not merely restricted to continuous spatio-temporal trajectory correspondence. This is known as the problem of person re-identification. Solving this problem has compelling practical applications, such as monitoring long-term activity patterns of targeted individuals over large spaces and across time zones. Current models mostly achieve inter-camera association by matching appearance and motion trends of a target across

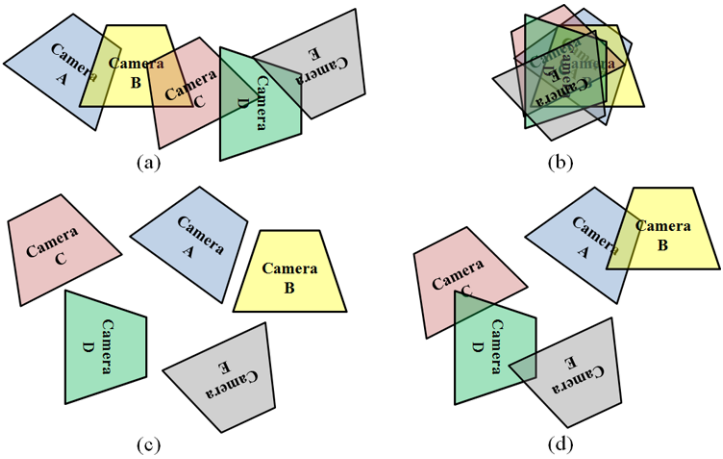


Fig. 2.11 An illustration of different degrees of overlap between the field of view of five cameras: (**a**) cameras partially overlap with adjacent cameras, (**b**) all cameras partially overlap with each other, (**c**) non-overlapping camera network, (**d**) the most common case involving different types of overlapping. The gap between two camera views is commonly known as a '*blind area*', in which a target object may be temporarily out of view of any camera

views. Appearance features are often extracted from the entire individual body since biometric features, such as iris or facial patterns, are no longer reliable in an unconstrained visual environment. Motion trends are typically represented by movement speed across views. Different strategies have been proposed to ensure those features to be robust against changes in object articulation, camera viewpoint, and illumination (Farenzena et al. 2010; Gray and Tao 2008; Hu et al. 2006; Prosser et al. 2008).

Object association based on visual appearance information alone is extremely difficult due to the visual ambiguities of people's appearance in an unconstrained public space (see Figs. 2.8 and 2.12). Apart from exploiting local context such as the group association context (Zheng et al. 2009), additional inter-camera spatio-temporal contextual information needs be explored to constrain and reduce association ambiguities across views. In general, inter-camera person re-identification cannot be achieved by matching imagery information alone. It is essential to formulate and model knowledge about inter-camera relationships as contextual constraints in assisting object re-identification over significantly different camera views (Chen et al. 2008; Loy et al. 2010).

If correspondences of many objects are established using a large pool of observations over time, a model can be built to infer the paths and transition time distributions between cameras. Makris et al. (2004) describe an unsupervised method to accumulate observations from a large set of cross-camera entrance and exit events in order to estimate an inter-camera transition time distribution. A peak in this transition time distribution implies a connection between the two camera views. Alternatively, without having to solve any object motion displacement correspondence

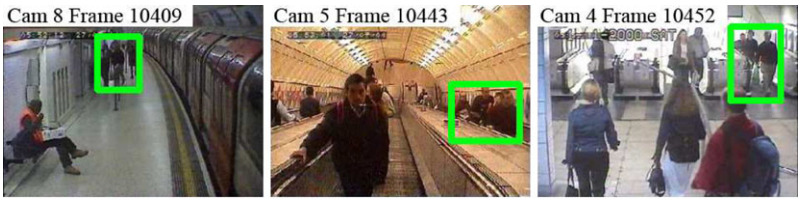


Fig. 2.12 Examples of significant changes in the visual appearance of people moving across different camera views in a public space: (**a**) A group of people (highlighted in *green boxes*) get off a train [Cam 8, frame 10409], and subsequently (**b**) take an upward escalator [Cam 5, frame 10443], which (**c**) leads them to the escalator exit view [Cam 4, frame 10452]. The same people exhibit drastic appearance variations due to changes in illumination (both intra and inter-camera), camera viewpoint, and each individual's body articulation

problem explicitly, another strategy applicable to constrain object association across disjoint cameras is to infer inter-camera relationships through modelling stochastic temporal correlation patterns across views from a population of activities instead of any individual's whereabouts (Loy et al. 2010; Marinakis and Dudek 2009; Niu and Grimson 2006; Tieu et al. 2005).

Estimation of inter-camera behaviour correlation can be made easier if the topology of a camera network can be inferred. This problem of inferring the spatial and temporal relationships among cameras is known as camera topology inference (Makris et al. 2004; Tieu et al. 2005). The aim is to estimate, given population activity patterns across camera views, camera transition probabilities, i.e. how likely an object exiting a camera view would reappear in another camera view; and inter-camera time delay distributions, which is the travel time needed to cross a blind area. The resulting camera topology can be represented by a graph, whose nodes describing population activities in camera field of views, whilst connections between nodes depict either camera transition probabilities or inter-camera time delay distributions.

2.8 Holistic Awareness: Connecting the Dots

Global activity analysis across multiple camera views and multiple information sources, as if 'connecting the dots', is an emerging problem that demands a solution. The goal is to build a global model for connecting and interpreting holistically distributed activities captured locally by different cameras and sensors, for example, detecting unusual events with a global situational awareness. In a distributed camera network, a global interpretation of localised activities is subject to uncertain and partial observations at different locations and different time, compounded by significant differences in object visual appearance (Fig. 2.12).

An intuitive approach to global activity analysis is to reconstruct a global trail of an object by merging the trajectories observed in different views (Zelniker et al. 2008). With non-overlapping camera views, one also needs to solve the camera

topology inference problem (Makris et al. 2004; Tieu et al. 2005) and the object association problem (Javed et al. 2005). However, in a crowded public space, it is unrealistic to visually reconstruct exhaustively object global motion trajectories for all objects.

Wang et al. (2010) suggest a trajectory-based solution which bypasses both the topology inference and object association problems. In particular, trajectories observed in different views are first connected based on their temporal proximity, that is, two trajectories are connected if they approximately co-occur in time. A generative model based on latent Dirichlet allocation (Blei et al. 2003) is exploited to cluster object trajectories by co-occurrence. The formed clusters are used to represent activity categories for imposing expectation on inter-camera transition probabilities. A reconstructed global trajectory is considered 'unusual' if it has a low likelihood when matched against the inter-camera transition probabilities.

In contrast to a trajectory-based model, Loy et al. (2010) consider an alternative model that automatically infers the unknown time-delayed correlations among local activities across cameras without object segmentation and tracking. This model aims to explore visual context in a global sense. Human behaviours in a public space are inherently context-aware, and context-constrained by implicit or explicit physical structures and social protocols observed in a shared space. Automatically discovering global spatio-temporal correlations among different camera locations over time can help to establish more accurate interpretations of object behaviours not only visually plausible at each camera location but also globally coherent in context.

Most multi-camera systems focus on passive and stationary cameras. There is a growing interest in collaborative activity analysis using a combination of heterogeneous sensors including PTZ cameras, mobile cameras, non-visual positioning sensors, and radio tags. There are systems using different types of cameras for collaborative monitoring, for example, a static camera to monitor a wide-area scene, working with a PTZ camera to obtain closed-up shot of a facial image (Zhou et al. 2003). More recently, there are also studies that investigate cooperative object detection and tracking using multiple active PTZ cameras (Everts et al. 2007) and through wearable cameras (Alahia et al. 2010). However, the use of heterogeneous sensors remains largely unexplored, especially on the development of reliable models for harvesting and fusing multi-channel information from different types of sensors and non-sensor sources.

References

Aghajanian, J., Warrell, J., Prince, S., Li, P., Rohn, J., Baum, B.: Patch-based within-object classification. In: IEEE International Conference on Computer Vision, Kyoto, Japan, October 2009, pp. 1125–1132 (2009)

Alahia, A., Vandergheynsta, P., Bierlaireb, M., Kunt, M.: Cascade of descriptors to detect and track objects across any network of cameras. Comput. Vis. Image Underst. **114**(6), 624–640 (2010)

Ali, A., Aggarwal, J.: Segmentation and recognition of continuous human activity. In: IEEE Workshop on Detection and Recognition of Events in Video, Vancouver, Canada, July 2001, pp. 28–35 (2001)

Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: European Conference on Computer Vision, Marseille, France, October 2008, pp. 1–14 (2008)

Ambady, N., Rosenthal, R.: Thin slices of expressive behaviour as predictors of interpersonal consequences: a meta-analysis. Psychol. Bull. **111**(2), 256–274 (1992)

Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: people detection and articulated pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, June 2009, pp. 1014–1021 (2009)

Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K., Kollias, S.: Emotion analysis in man-machine interaction systems. In: Machine Learning for Multimodal Interaction, pp. 318–328 (2005)

Bar, M.: Visual objects in context. Nat. Rev., Neurosci. **5**, 617–629 (2004)

Bar, M., Aminoff, E.: Cortical analysis of visual context. Neuron **38**(2), 347–358 (2003)

Bar, M., Ullman, S.: Spatial context in recognition. Perception **25**, 343–352 (1993)

Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA, June 2005, pp. 568–573 (2005)

Bashir, K., Xiang, T., Gong, S.: Gait recognition without subject cooperation. Pattern Recognit. Lett. **31**(13), 2052–2060 (2010)

Bassili, J.N.: Emotion recognition: the role of facial movement and the relative importance of upper and lower area of the face. J. Pers. Soc. Psychol. **37**(11), 2049–2058 (1979)

Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002)

Benfold, B., Reid, I.: Colour invariant head pose classification in low resolution video. In: British Machine Vision Conference, Leeds, UK, September 2008

Benfold, B., Reid, I.: Guiding visual surveillance by tracking human attention. In: British Machine Vision Conference, London, UK, September 2009

Biederman, I., Mezzanotte, R.J., Rabinowitz, J.C.: Scene perception: detecting and judging objects undergoing relational violations. Cogn. Psychol. **14**, 143–177 (1982)

Black, M., Jepson, A.: Recognising temporal trajectories using the condensation algorithm. In: IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, pp. 16–21 (1998)

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Bobick, A.F., Davis, J.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)

Bobick, A.F., Wilson, A.: A state-based technique for the summarisation of recognition of gesture. In: IEEE International Conference on Computer Vision, pp. 382–388 (1995)

Bregonzio, M., Gong, S., Xiang, T.: Recognising action as clouds of space-time interest points. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, June 2009, pp. 1948–1955 (2009)

Breitenstein, M.D.: Visual surveillance—dynamic behavior analysis at multiple levels. PhD thesis, ETH Zurich (2009)

Breitenstein, M.D., Sommerlade, E., Leibe, B., Van Gool, L., Reid, I.: Probabilistic parameter selection for learning scene structure from video. In: British Machine Vision Conference, Leeds, UK, September 2008

Burgoon, J., Adkins, M., Kruse, J., Jensen, M.L., Meservy, T., Twitchell, D.P., Deokar, A., Nunamaker, J.F., Lu, S., Tsechpenakis, G., Metaxas, D.N., Younger, R.E.: An approach for intent identification by building on deception detection. In: 38th Annual Hawaii International Conference on System Sciences, Big Island, Hawaii, USA, January 2005, pp. 21–28 (2005a)

Burgoon, J., Jensen, M.L., Meservy, T., Kruse, J., Nunamaker, J.F.: Augmenting human identification of emotional states in video. In: International Conference on Intelligent Data Analysis, McClean, USA (2005b)

Chang, Y., Hu, C., Turk, M.: Probabilistic expression analysis on manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition, Washington DC, USA, June 2004, pp. 520–527 (2004)

Chen, K., Lai, C., Hung, Y., Chen, C.: An adaptive learning method for target tracking across multiple cameras. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)

Cohen, I., Sebe, N., Garg, A., Chen, L., Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling. Comput. Vis. Image Underst. **91**, 160–187 (2003)

Cohen, N., Gatusso, J., MacLennan-Brown, K.: CCTV Operational Requirements Manual—Is Your CCTV System Fit for Purpose? Home Office Scientific Development Branch, version 4 (55/06) edition (2006)

Coulson, M.: Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. J. Nonverbal Behav. **28**(2), 117–139 (2004)

Darwin, C.: The Expression of the Emotions in Man and Animals. Murray, London (1872)

Davis, J., Bobick, A.F.: The representation and recognition of action using temporal templates. In: IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico, June 1997, pp. 928–934 (1997)

Donato, G., Bartlett, M., Hager, J., Ekman, P., Sejnowski, T.: Classifying facial actions. IEEE Trans. Pattern Anal. Mach. Intell. **21**(10), 974–989 (1999)

Dornaika, F., Davoine, F.: Simultaneous facial action tracking and expression recognition using a particle filter. In: IEEE International Conference on Computer Vision, Beijing, China, October 2005, pp. 1733–1738 (2005)

Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision, Nice, France, pp. 726–733 (2003)

Ekman, P., Friesen, W.V.: Pictures of Facial Affect. Consulting Psychologists Press, Palo Alto (1976)

Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)

Essa, I., Pentland, A.: Coding, analysis, interpretation, and recognition of facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 757–763 (1997)

Everts, I., Sebe, N., Jones, G.A.: Cooperative object tracking with multiple PTZ cameras. In: International Conference on Image Analysis and Processing, September 2007, pp. 323–330 (2007)

Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, June 2010, pp. 2360–2367 (2010)

Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognit. **36**, 259–275 (2003)

Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. Int. J. Comput. Vis. **61**(1), 55–79 (2005)

Fernández, C., Gonzàlez, J., Roca, X.: Automatic learning of background semantics in generic surveilled scenes. In: European Conference on Computer Vision, Crete, Greece, October 2010, pp. 678–692 (2010)

Freeman, W.T., Tanaka, K., Ohta, J., Kyuma, K.: Computer vision for computer games. In: IEEE International Conference on Automatic Face and Gesture Recognition, Killington, USA, October 1996, pp. 100–105 (1996)

Gilbert, A., Illingworth, J., Bowden, R.: Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In: European Conference on Computer Vision, pp. 222–233 (2008)

Gong, S., Buxton, H.: Bayesian nets for mapping contextual knowledge to computational constraints in motion segmentation and tracking. In: British Machine Vision Conference, Guildford, UK, September 1993, pp. 229–238 (1993)

Gong, S., McKenna, S., Collins, J.J.: An investigation into face pose distributions. In: IEEE International Conference on Automatic Face and Gesture Recognition, Killington, USA, October 1996, pp. 265–270 (1996)

Gong, S., Walter, M., Psarrou, A.: Recognition of temporal structures: learning prior and propagating observation augmented densities via hidden Markov states. In: IEEE International Conference on Computer Vision, Corfu, Greece, September 1999, pp. 157–162 (1999)

Gong, S., Xiang, T., Hongeng, S.: Learning human pose in crowd. In: Workshop on Multimodal Pervasive Video Analysis, ACM International Conference on Multimedia, Firenze, Italy, October 2010

Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 2247–2253 (2007)

Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision, pp. 262–275 (2008)

Gunes, H., Piccardi, M.: A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: International Conference on Pattern Recognition, pp. 1148–1153 (2006)

Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. J. Netw. Comput. Appl. **30**(4), 1334–1345 (2007)

Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Trans. Pattern Anal. Mach. Intell. **28**(2), 316–322 (2006)

Hoey, J., Little, J.J.: Value directed learning of gestures and facial displays. In: IEEE Conference on Computer Vision and Pattern Recognition, Washington DC, USA, June 2004, pp. 1026–1033 (2004)

Hospedales, T., Gong, S., Xiang, T.: A Markov clustering topic model for mining behaviour in video. In: IEEE International Conference on Computer Vision, Kyoto, Japan, October 2009, pp. 1165–1172 (2009)

Hu, W., Hu, M., Zhou, X., Tan, T., Lou, J., Maybank, S.: Principal axis-based correspondence between multiple cameras for people tracking. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 663–671 (2006)

Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: IEEE International Conference on Computer Vision, Kyoto, Japan, October 2009, pp. 128–135 (2009)

Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking across multiple cameras with disjoint views. In: IEEE International Conference on Computer Vision, pp. 952–957 (2003)

Javed, O., Shafique, K., Shah, M.: Appearance modeling for tracking in multiple non-overlapping cameras. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 26–33 (2005)

Jia, K., Gong, S.: Multi-resolution patch tensor for facial expression hallucination. In: IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, June 2006, pp. 395–402 (2006)
Jia, K., Gong, S.: Generalised face super-resolution. IEEE Trans. Image Process. **17**(6), 873–886 (2008)
Jiang, H.: Human pose estimation using consistent max-covering. In: IEEE International Conference on Computer Vision, Kyoto, Japan, October 2009, pp. 1357–1364 (2009)
Kaliouby, R.E., Robinson, P.: Real-time inference of complex mental states from facial expressions and head gestures. In: IEEE Workshop on Real-Time Vision for Human-Computer Interaction, Washington DC, USA, June 2004
Kapoor, A., Picard, R.W.: Multimodal affect recognition in learning environments. In: ACM International Conference on Multimedia, pp. 677–682 (2005)
Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: IEEE International Conference on Computer Vision, pp. 166–173 (2005)
Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: IEEE International Conference on Computer Vision, Rio de Janeiro, Brasil, October 2007, pp. 1–8 (2007)
Kim, J., Grauman, K.: Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, June 2009, pp. 2921–2928 (2009)
Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1446–1453 (2009)
Kruegle, H.: CCTV Surveillance: Video Practices and Technology. Butterworth/Heinemann, Stoneham/Berlin (2006)
Kuettel, D., Breitenstein, M.D., Van Gool, L., Ferrari, V.: What's going on? discovering spatio-temporal dependencies in dynamic scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, June 2010, pp. 1951–1958 (2010)
Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**(2), 107–123 (2005)
Laptev, I., Lindeberg, T.: Space-time interest points. In: IEEE International Conference on Computer Vision, pp. 432–439 (2003)
Lee, C.S., Elgammal, A.: Facial expression analysis using nonlinear decomposable generative models. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, pp. 17–31 (2005)
Lee, M.W., Cohen, I.: Human upper body pose estimation in static images. In: European Conference on Computer Vision, Cambridge, UK, pp. 126–138 (2004)
Li, J., Gong, S., Xiang, T.: Global behaviour inference using probabilistic latent semantic analysis. In: British Machine Vision Conference, Leeds, UK, pp. 193–202 (2008a)
Li, J., Gong, S., Xiang, T.: Scene segmentation for behaviour correlation. In: European Conference on Computer Vision, Marseille, France, pp. 383–395 (2008b)
Liu, J., Shah, M.: Learning human actions via information maximization. In: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, June 2008, pp. 1–8 (2008)
Loy, C.C., Xiang, T., Gong, S.: Time-delayed correlation analysis for multi-camera activity understanding. Int. J. Comput. Vis. **90**(1), 106–129 (2010)
Loy, C.C., Xiang, T., Gong, S.: Detecting and discriminating behavioural anomalies. Pattern Recognit. **44**(1), 117–132 (2011)
Lyons, M.J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. IEEE Trans. Pattern Anal. Mach. Intell. **21**(12), 1357–1362 (1999)
Makris, D., Ellis, T.: Learning semantic scene models from observing activity in visual surveillance. IEEE Trans. Syst. Man Cybern. **35**(3), 397–408 (2005)

Makris, D., Ellis, T., Black, J.: Bridging the gaps between cameras. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 205–210 (2004)
Marinakis, D., Dudek, G.: Self-calibration of a vision-based sensor network. Image Vis. Comput. **27**(1–2), 116–130 (2009)
McKenna, S., Gong, S.: Gesture recognition for visually mediated interaction using probabilistic event trajectories. In: British Machine Vision Conference, Southampton, UK, pp. 498–508 (1998)
Meeren, H., Heijnsbergen, C., Gelder, B.: Rapid perceptual integration of facial expression and emotional body language. Proc. Natl. Acad. Sci. USA **102**(45), 16518–16523 (2005)
Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behaviour detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 935–942 (2009)
Ng, J., Gong, S.: Composite support vector machines for the detection of faces across views and pose estimation. Image Vis. Comput. **20**(5–6), 359–368 (2002)
Niu, C., Grimson, W.E.L.: Recovering non-overlapping network topology using far-field vehicle tracking data. In: International Conference of Pattern Recognition, pp. 944–949 (2006)
Nixon, M.S., Tan, T., Chellappa, R.: Human Identification Based on Gait. Springer, Berlin (2005)
Okuma, K., Taleghani, A., Freitas, N.D., Little, J., Lowe, D.: A boosted particle filter: multitarget detection and tracking. In: European Conference on Computer Vision (2004)
Orozco, J., Gong, S., Xiang, T.: Head pose classification in crowded scenes. In: British Machine Vision Conference, London, UK, September 2009
Ozuysal, M., Lepetit, V., Fua, P.: Pose estimation for category specific multiview object localization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 778–785 (2009)
Palmer, S.: The effects of contextual scenes on the identification of objects. Mem. Cogn. **3**, 519–526 (1975)
Pantic, M., Patras, I.: Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans. Syst. Man Cybern. **36**(2), 433–449 (2006)
Pantic, M., Rothkrantz, L.: Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1424–1445 (2000)
Pantic, M., Rothkrantz, L.: Toward an affect-sensitive multimodal human-computer interaction. In: Proceeding of the IEEE, vol. 91, pp. 1370–1390 (2003)
Pantic, M., Rothkrantz, L.: Facial action recognition for facial expression analysis from static face images. IEEE Trans. Syst. Man Cybern. **34**(3), 1449–1461 (2004)
Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: IEEE International Conference on Multimedia and Expo, pp. 317–321 (2005)
Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: a review. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 677–695 (1997)
Prosser, B., Gong, S., Xiang, T.: Multi-camera matching using bi-directional cumulative brightness transfer functions. In: British Machine Vision Conference, Leeds, UK, September 2008
Psarrou, A., Gong, S., Buxton, H.: Modelling spatio-temporal trajectories and face signatures on partially recurrent neural networks. In: IEEE International Joint Conference on Neural Networks, Perth, Australia, pp. 2226–3321 (1995)
Psarrou, A., Gong, S., Walter, M.: Recognition of human gestures and behaviour based on motion trajectories. Image Vis. Comput. **20**(5–6), 349–358 (2002)

Ramanan, D., Forsyth, D.A.: Automatic annotation of everyday movements. In: Advances in Neural Information Processing Systems, Vancouver, Canada (2003)

Rao, C., Shah, M.: View-invariance in action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 316–322 (2001)

Ravindra De Silva, P., Bianchi-Berthouze, N.: Modeling human affective postures: an information theoretic characterisation of posture features. Comput. Animat. Virtual Worlds **15**, 169–276 (2004)

Ravindra De Silva, P., Osano, M., Marasinghe, A.: Towards recognizing emotion with affective dimensions through body gestures. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 269–274 (2006)

Robertson, N., Reid, I.: Estimating gaze direction from low-resolution faces in video. In: European Conference on Computer Vision, pp. 402–415 (2006)

Schlenzig, J., Hunter, E., Jain, R.: Recursive identification of gesture inputs using hidden Markov model. In: Workshop on Applications of Computer Vision, pp. 187–194 (1994)

Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: International Conference on Pattern Recognition, Cambridge, UK, pp. 32–36 (2004)

Schwartz, O., Hsu, A., Dayan, P.: Space and time in visual context. Nat. Rev., Neurosci. **8**, 522–535 (2007)

Shan, C., Gong, S., McOwan, P.: Appearance manifold of facial expression. In: Sebe, N., Lew, M.S., Huang, T.S. (eds.) Computer Vision in Human-Computer Interaction. Lecture Notes in Computer Science, vol. 3723, pp. 221–230. Springer, Berlin (2005)

Shan, C., Gong, S., McOwan, P.: Capturing correlations among facial parts for facial expression analysis. In: British Machine Vision Conference, Warwick, UK, September 2007

Shan, C., Gong, S., McOwan, P.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)

Sheikh, Y., Sheikh, M., Shah, M.: Exploring the space of a human action. In: IEEE International Conference on Computer Vision, Beijing, China, October 2005, pp. 144–149 (2005)

Sherrah, J., Gong, S.: Fusion of perceptual cues for robust tracking of head pose and position. Pattern Recognit. **34**(8), 1565–1572 (2001)

Sherrah, J., Gong, S., Ong, E.-J.: Face distribution in similarity space under varying head pose. Image Vis. Comput. **19**(11), 807–819 (2001)

Siva, P., Xiang, T.: Action detection in crowd. In: British Machine Vision Conference, Aberystwyth, UK, September 2010

Starner, T., Pentland, A.: Visual recognition of American sign language using hidden Markov models. In: IEEE International Conference on Automatic Face and Gesture Recognition, Zurich, Switzerland, pp. 189–194 (1995)

Stauffer, C., Tieu, K.: Automated multi-camera planar tracking correspondence modeling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 259–266 (2003)

Streib, K., Davis, J.W.: Extracting pathlets from weak tracking data. In: IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, USA, pp. 353–360 (2010)

Suwa, M., Sugie, N., Fujimora, K.: A preliminary note on pattern recognition of human emotional expression. In: International Joint Conference on Pattern Recognition, pp. 408–410 (1978)

Tian, Y.: Evaluation of face resolution for expression analysis. In: Computer Vision and Pattern Recognition Workshop, pp. 82–89 (2004)

Tian, Y., Kanade, T., Cohn, J.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2), 97–115 (2001)

Tieu, K., Dalley, G., Grimson, W.E.L.: Inference of non-overlapping camera network topology by measuring statistical dependence. In: IEEE International Conference on Computer Vision, pp. 1842–1849 (2005)

Valstar, M., Pantic, M.: Fully automatic facial action unit detection and temporal analysis. In: Computer Vision and Pattern Recognition Workshop, pp. 149–154 (2006)

Valstar, M., Patras, I., Pantic, M.: Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In: Computer Vision and Pattern Recognition Workshop, vol. 3, pp. 76–84 (2005)

Van Gool, L., Breitenstein, M., Gammeter, S., Grabner, H., Quack, T.: Mining from large image sets. In: ACM International Conference on Image and Video Retrieval, Santorini, Greece, pp. 10:1–10:8 (2009)

Vrij, A.: Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice. Wiley, New York (2000)

Vrij, A., Mann, S.: Police use of nonverbal behaviour as indicators of deception. In: Riggio, R.E., Feldman, R.S. (eds.) Applications of Nonverbal Communication, pp. 63–94. Erlbaum, Hillsdale (2005)

Wang, S., Quattoni, A., Morency, L.-P., Demirdjian, D., Darrell, T.: Hidden conditional random fields for gesture recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1521–1527 (2006a)

Wang, X., Tieu, K., Grimson, W.E.L.: Learning semantic scene models by trajectory analysis. In: European Conference on Computer Vision, pp. 110–123 (2006b)

Wang, X., Ma, X., Grimson, W.E.L.: Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 539–555 (2009)

Wang, X., Tieu, K., Grimson, W.E.L.: Correspondence-free activity analysis and scene modeling in multiple camera views. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 56–71 (2010)

Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: real-time tracking of the human body. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 780–785 (1997)

Wu, B., Nevatia, R.: Detection and tracking of multiple partially occluded humans by Bayesian combination of edgelet based part detectors. Int. J. Comput. Vis. **75**(2), 247–266 (2007)

Wu, Y., Huang, T.S.: Human hand modeling, analysis and animation in the context of human computer interaction. IEEE Signal Process. Mag. **18**(3), 51–60 (2001)

Yacoob, Y., Davis, L.S.: Recognizing human facial expression from long image sequences using optical flow. IEEE Trans. Pattern Anal. Mach. Intell. **18**(6), 636–642 (1996)

Yang, M., Wu, Y., Hua, G.: Context-aware visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. **31**(7), 1195–1209 (2008)

Yang, W., Wang, Y., Mori, G.: Efficient human action detection using a transferable distance function. In: Asian Conference on Computer Vision, pp. 417–426 (2009)

Yeasin, M., Bullot, B., Sharma, R.: From facial expression to level of interests: a spatio-temporal approach. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 922–927 (2004)

Yilmaz, A., Shah, M.: Actions sketch: a novel action representation. In: IEEE Conference on Computer Vision and Pattern Recognition, Washington DC, USA, pp. 984–989 (2005)

Yin, L., Loi, J., Xiong, W.: Facial expression representation and recognition based on texture augmentation and topographic masking. In: ACM Multimedia, New York, USA, pp. 236–239 (2004)

Yuan, J.S., Liu, Z.C., Wu, Y.: Discriminative subvolume search for efficient action detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2442–2449 (2009)

Zaharescu, A., Wildes, R.: Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In: European Conference on Computer Vision, Crete, Greece, October 2010, pp. 563–576 (2010)

Zalewski, L., Gong, S.: 2D statistical models of facial expressions for realistic 3D avatar animation. In: IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA, June 2005, pp. 217–222 (2005)

Zelniker, E.E., Gong, S., Xiang, T.: Global abnormal behaviour detection using a network of CCTV cameras. In: IEEE International Workshop on Visual Surveillance, Marseille, France, October 2008

Zhang, J., Gong, S.: Action categorisation by structural probabilistic latent semantic analysis. Comput. Vis. Image Underst. **114**(8), 857–864 (2010a)

Zhang, J., Gong, S.: Action categorisation with modified hidden conditional random field. Pattern Recognit. **43**(1), 197–203 (2010b)

Zhang, J., Zhou, S., McMillan, L., Comaniciu, D.: Joint real-time object detection and pose estimation using probabilistic boosting network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2007)

Zhang, Y., Ji, Q.: Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5), 1–16 (2005)

Zhang, Z., Lyons, M.J., Schuster, M., Akamatsu, S.: Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron. In: IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, April 1998, pp. 454–459 (1998)

Zhang, Z., Hu, Y., Chan, S., Chia, L.: Motion context: a new representation for human action recognition. In: European Conference on Computer Vision, pp. 817–829 (2008)

Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: British Machine Vision Conference, London, UK, September 2009

Zhou, X., Collins, R., Kanade, T.: A master-slave system to acquire biometric imagery of humans at distance. In: International Workshop on Video Surveillance, pp. 113–120 (2003)



http://www.springer.com/978-0-85729-669-6

Visual Analysis of Behaviour
From Pixels to Semantics
Gong, S.; Xiang, T.
2011, XIX, 356 p., Hardcover
ISBN: 978-0-85729-669-6