HUMAN POSE AND ACTION RECOGNITION USING NEGATIVE SPACE ANALYSIS

Michaella Janse van Vuuren

A Thesis presented for the degree of
DOCTOR OF PHILOSOPHY
in the Department of Electrical Engineering
UNIVERSITY OF CAPE TOWN
August 2004

HUMAN POSE AND ACTION RECOGNITION USING NEGATIVE SPACE ANALYSIS

Michaella Janse van Vuuren

August 2004

Abstract This thesis proposes a novel approach to extracting pose information from image sequences. Current state of the art techniques focus exclusively on the image space occupied by the body for pose and action recognition. The method proposed here, however, focuses on the negative spaces: the areas surrounding the individual. This has resulted in the colour-coded negative space approach, an image preprocessing step that circumvents the need for complicated model fitting or template matching methods. The approach can be described as follows: negative spaces surrounding the human silhouette are extracted using horizontal and vertical scanning processes. These negative space areas are more numerous, and undergo more radical changes in shape than the single area occupied by the figure of the person performing an action. The colour-coded negative space representation is formed using the four binary images produced by the scanning processes. Features are then extracted from the colour-coded images. These are based on the percentage of area occupied by distinct coloured regions as well as the bounding box proportions. Pose clusters are identified using feedback from an independant action set. Subsequent images are classified using a simple Euclidean distance measure. An image sequence is thus temporally segmented into its corresponding pose representations. Action recognition simply becomes the detection of a temporally ordered sequence of poses that characterises the action. The method is purely vision-based, utilising monocular images with no need for body markers or special clothing.

Two datasets were constructed using several actors performing different poses and actions. Some of these actions included actors waving their arms, sitting down or kicking a leg. These actions were recorded against a monochrome background to simplify the segmentation of the actors from the background. The actions were then recorded on DV

cam and digitised into a data base. The silhouette images from these actions were isolated and placed in a frame or bounding box. The next step was to highlight the negative spaces using a directional scanning method. This scanning method colour-codes the negative spaces of each action. What became immediately apparent is that very distinctive colour patterns formed for different actions. To emphasise the action, different colours were allocated to negative spaces surrounding the image. For example, the space between the legs of an actor standing in a T - pose with legs apart would be allocated yellow, while the space below the arms were allocated different shades of green. The space surrounding the head would be different shades of purple. During an action when the actor moves one leg up in a kicking fashion, the yellow colour would increase. Inversely, when the actor closes his legs and puts them together, the yellow colour filling the negative space would decrease substantially. What also became apparent is that these coloured negative spaces are interdependent and that they influence each other during the course of an action. For example, when an actor lifts one of his legs, increasing the yellow-coded negative space, the green space between that leg and the arm decreases. This interrelationship between colours hold true for all poses and actions as presented in this thesis. In terms of pose recognition, it is significant that these colour coded negative spaces and the way the change during an action or a movement are substantial and instantly recognisable. Compare for example, looking at someone lifting an arm as opposed to seeing a vast negative space changing shape. In a controlled research environment, several actors were instructed to perform a number of different actions. After colour coding the negative spaces, it became apparent that every action can be recognised by a unique colour coded pattern.

The challenge is to ascribe a numerical presentation, a mathematical quotation, to extract the essence of what is so visually apparent. The essence of pose recognition and it's measurability lies in the relationship between the colours in these negative spaces and how they impact on each other during a pose or an action. The simplest way of measuring this relationship is by calculating the percentage of each colour present during an action. These calculated percentages become the basis of pose and action recognition. By plotting these percentages on a graph confirms that the essence of these different actions and poses can in fact been captured and recognised. Despite variations in these traces caused by time

differences, personal appearance and mannerisms, what emerged is a clear recognisable pattern that can be married to an action or different parts of an action. 7 Actors might lift their left leg, some slightly higher than others, some slower than others and these variations in terms of colour percentages would be recorded as a trace, but there would be very specific stages during the action where the traces would correspond, making the action recognisable.

In conclusion, using negative space as a tool in human pose and tracking recognition presents an exiting research avenue because it is influenced less by variations such as difference in personal appearance and changes in the angle of observation. This approach is also simplistic and does not rely on complicated models and templates

Acknowledgements

Thanks to the people in my research group especially Frances, Arne, Jerome, and my supervisor Professor Gerhard de Jager. To Frans Lewis without whom I would not have been able to obtain my data. To all my family, pets and friends, especially Eugene, Linden and Albert, for their unwavering support. To Greg and Chris for their support and de Beers funding. And lastly for the National Research Fund for their financial contribution to this work.

Contents

List of Figures

Chapter 1

Introduction

1.1 Research Context

The bombing of the World Trade centre has brought to new focus the importance of human surveillance systems. Previously, there was a concern that these systems would invade people's privacy, but now the concern weighs more toward protecting people against potential threats. Smart rooms and buildings using high tech surveillance systems, refer to architectural spaces that respond to the individuals in them [62]. These environments can track people and monitor their activities, identifying patterns of behaviour that can be used to assist individuals or detect suspicious behaviour. Whether the system is used for identifying harmful behaviour or assisting in every day tasks the basic requirements of intelligent surveillance are the same.

To implement a human action or gesture recognition system, the actions of the individuals need to be captured on video and then analysed by a computer system. In most cases the human form needs to be separated from the background. Features, such as edges, colour, or motion are extracted from the segmented images. These features can be used to characterise or model the action or pose contained within the images for further recognition. The pose or action information can be used for surveillance, teleconferencing, games, interactive virtual worlds, character animation, sign language translation, gesture driven control, biometrics, video content based indexing and choreography.

Researchers developing human recognition systems have focused on extracting features

from the space occupied by the body of the person or the entire image [54] to extract information or build models. What distinguishes the work presented here is a reliance rather on the negative space to recognise poses and actions. Information about a pose is contained in the regions surrounding the body, the negative spaces, these areas form patterns when the body moves that uniquely describes the pose and by implication the action contained in a sequence. The premise of this thesis is that features extracted from this space can be used to develop a computer vision system that recognises human pose and action.

1.2 Objectives

The purpose of this dissertation is to show that analysis of the negative spaces can be used to recognise human poses and actions. The following assumptions and objectives have been made to achieve this goal.

The objectives are:

- To review current approaches to pose and action recognition.
- To construct two action sequence databases, one representing appearance variation and the other variation in the viewing angle.
- To base the recognition system on the negative space, no features relating to dimensional information about the space, or people, will be incorporated. This implies the system will not be biased by these features, and pose or action recognition results can be attributed solely to features derived from the negative space.

The assumptions:

- The system will attempt to recognise the actions of a single human, no other objects or interactions are considered.
- Segmentation is simplified using Chroma-key, or blue screen techniques.

- Loose fitting clothes can be worn, no markers or special clothing is used. The only constraints placed on the attire of the actors is that the colour should differ from the blue Chroma-key background, and that dresses were not included.
- No camera calibration is used.
- Self occlusions are allowed.
- No pose initialisation is used.

1.3 Outline

The dissertation introduces a vision based method that recognises human poses and actions using features extracted from the negative spaces surrounding the individual. This section provides a short outline of how the document relates the discovery.

The literature review in Chapter 2 describes the various methods that researchers have employed to solve the problem of human tracking, pose and action recognition.

The central theme of this dissertation is negative space, as there is no relevant recognition literature that utilises this image space, a section is included that provides background on the theory of negative space from an art perspective.

Chapter 3 relates how the datasets used in the thesis was acquired and processed. The data was captured from two different studios, one containing a single camera and another where it was possible to film an action from six cameras angles simultaneously. Both these studios use blue Chroma-key backgrounds to simplify the segmentation of a person from the background. The multi-view dataset contains actions filmed from six different angles, each separated by 30 degrees. The actions are performed eight times by seven different individuals.

Chapter 4 describes how the segmented human silhouette is processed to obtain the colour coded representation. A bounding box is placed around the silhouette. This restricts the algorithm to the negative spaces between the bounding box and the silhouette figure. A scanning procedure is used to isolate these negative spaces and colour code the different

negative space areas. This very simple procedure transforms the silhouettes into coloured patterns that emphasise the differences in body poses. To recognise the poses contained in the images, characteristic features are extracted for further processing. These features are a description of the bounding box proportions and the percentage of each of the colours that are present in the image. The chapter contains a number of plots from both datasets that show how the relationships between the features change for a number of actions. There is also a discussion on how difficult it is to recognise entire action traces due to the different ways people perform the same action.

Chapter 5 describes how the negative space features are used, in conjunction with a feedback action set, to construct a pose classifier. The pose classifier makes it possible to deconstruct action sequences into a pose labelled representation. Action recognition becomes the detection of a unique sequence of these pose labels in the input.

Chapter 6 discusses the results of the pose and action recognition systems and Chapter 7 concludes the document.

Human action recognition is a dynamic process. The methods and discoveries described in this document therefore do not translate well into static imagery. A better understanding can be gained by viewing the accompanying Appendix CD. This multimedia document contains several animations and dynamic images that are described in the dissertation. It highly recommended that the CD be viewed before the dissertation is read.

Chapter 2

A Review of Human Pose and Action Recognition

This chapter reviews the relevant research done in the field of human pose, action and gesture recognition. This is followed by a discussion on negative space, an avenue of research that has not been investigated in terms of human action recognition.

The work involving the recognition of human movement can be divided into three broad areas of interest, human body pose estimation, person tracking and actual gesture or action recognition [1, 23]. The general framework used in this dissertation for describing the human motion capture follows along the lines proposed by Moeslund and Granum [54]. Human recognition systems can be described in the terms of the following processes: initialisation, tracking, pose estimation, and recognition. Generally a system needs to be initialised before it can process data, the person is then identified by some form of segmentation from the background. Then the pose is estimated and the action recognised.

Typical assumptions or restrictions in this field refer to either movement or appearance. Movement assumptions are dependent on both the subject and /or camera(s). Appearance assumptions refer to limits placed on a subject and its environment.

Assumptions regarding movement are: the subject remains inside the camera view, constant or no camera movement, one person at a time, the subject faces the camera all the time, movements are parallel to the camera plane, no occlusions, slow movements, only one or two limbs move, a known motion pattern, the subject moves on flat ground.

Assumptions regarding appearance are: constant lighting, static background, uniform background, known camera parameters, special hardware (IR etc). Constraints imposed on the person are: known start pose, known subject, markers on the person, special coloured clothes, tight fitting clothes.

Initialisation concerns mainly the model initialisation, camera calibration and adaptation to scene characteristics. Off-line initialisation is used to find segmentation thresholds, camera calibration and the capturing of reference images.

Some model-based systems require an initial estimation of the pose, this is done by either requiring a special start pose [17] or by using an operator to specify the start pose manually [64, 86]. Very few systems find the start pose automatically [52, 67, 71].

A model sometimes needs to be initialised for model-based approaches. Models can range in complexity from the average of a number of people [6] to a measure of the current subject [26, 44, 79]. Systems that create a personalised model rely on building a 3D shape [55, 59] or fitting the current data to a generic model [30, 86].

2.1 Tracking

Tracking is used to establish relationships between different parts of a person's body or relationships between consecutive frames. This information can be used to track moving "human like" objects over time [76], estimate the pose or for action recognition [65]. Pose estimation is done using low level information such as edges that are matched against a model [31] or high level information like the blobs that represent the hands and feet used in Pfinder [79].

To track the desired features they have to be segmented from the rest of the image. This is done using either temporal or spatial information. The use of temporal data assumes that the subject is the only moving object in the sequence. Temporal information can be used by subtracting images from the background [28, 58] or previous image/s [3, 43, 65], or by detecting the flow between reference features between image frames. This flow can be between points [83], features such as edges [25] or blobs [10].

The use of spatial data for segmentation implies either thresholding or statistical approaches. Thresholding is used when the colour of the person [7] and the background [18, 34] is sufficiently different to distinguish one easily from the other. The most widely used and robust segmentation method is Chroma-keying [54]. Chroma-keying methods place the subject in front of a matte coloured background, the person is segmented by detecting pixels that are outside the colour range of the background [73]. Markers that identify different body parts also fall under this category [12, 24]. These methods rely on a number of appearance assumptions hence their use in controlled environments.

Statistical approaches use characteristics such as colour or edges for segmentation and are more suited to unconstrained applications. A popular method is to find the mean and variance of the intensity or colour of each pixel position in a sequence of background images. The image is segmented by comparing and classifying the current image pixels as either belonging to the background or not based on the background statistics [82]. This method can be improved by removing subject shadows [50] or by using the blob-approach where the subject's statistics are modelled and images in the next frame are classified to either belong to the person or not [79].

Static or dynamic/active contours are statistical methods that track edge segments or other attributes. Static structures are predefined and represents a part of the person's outline or a part of the body. Active contours adjust to fit the image features, they can be used to extract the person's outline [6] or distinct body parts [42].

Previous sections described how segmented entities are obtained. These must then be represented in a more compact form to be processed further. Two representations, the object-based representation and the image-based representation are discussed. The object-based representation takes the output from the figure-ground segmentation and represents this data using points, boxes, silhouettes or blobs. Points are used to represent passive or active body markers [57, 72]. Bounding boxes containing the segmented regions can be tracked [58] or processed further to estimate the pose within these regions [79]. The silhouette representation can result from the segmentation process [28] or the tracking of an active contour [6], and can be tracked directly [28] or used for further pose processing [16, 27, 52]. In the blob representation the subject is represented as a collection of blobs with similar characteristics such as, coherent flow [41], similar colours [29] or

both [10].

In image-based representations pixels of the entire image or areas of interest are transformed to another coordinate space using transformations such as the Fourier, Principal Component Analysis, direct Cosine Transforms and Wavelets.

2.2 Pose estimation

To analyse or recognise human motion we ofter need to extract information from the image that will allow us to estimate the body pose of an individual. The estimated pose can be an integral part of the tracking algorithm or the output of the system. The desired pose information can range form general information on the pose to very precise information relating the the position of various limbs. To accomplish this goal researchers have used either a model-free, indirect model, or direct model approach.

Model-free methods incorporate no a priori human model but use points, simple shapes and stick figures to represent the pose. The points can be derived from markers attached to the body [40], or the three points representing the head and hands found by colour [53] or blob segmentation [79] can be used to find the body pose. Simple shapes like a boundary box [18] or ellipse [58] can be used as intermediate or final simple pose representation. Stick figures are uses to obtain an estimate of the human skeleton [7, 35]. Exemplars in the form of key-frames can be used characterise poses [13, 38, 48, 75, 84].

Indirect Models use information known beforehand to construct reference or look-up tables to help identify the pose [69]. A rough description of the body pose, or positions of the hands and head, are examples of the poses that are recognised. The aspect ratios between various limbs can be used [11, 28], as can edges matched to similar structures in the model [47]. The rough overall pose can be detected and then used to find the individual body parts [28]. Behaviour models can be used to predict the motion beforehand, so as to improve the pose estimation [34, 80]. A model can also be used to ensure the validity of a detected pose [32, 33, 61].

Direct Models use a detailed a-priori model that is continuously updated by observations. The pose of the person can thus be obtained at any time from the matched stored model.

The skeleton structure of the body can be modelled as a tree of rigid parts and connected joints. Knowledge of the motions that can be performed by these joints imposes constraints on the possible model configurations. Flesh on the skeletal structure can then be modelled using primitive volumes, such as elliptical cylinders centred on the skeleton. The models used for pose estimation can range from simple stick-figures to 2-D contour models and volumetric 3-D models.

The stick-figure [15], representation of a human, models the limbs and body parts as line segments linked by joints. The 3-D locations of joints or feature points are determined using this model.

Papers that describe the use of 2-D models include [6, 41, 45, 46, 85]. The work of Ju, Black and Yacoob [41] involve the development of a two- dimensional "cardboard person model". This model represents the relative positions and sizes of the body parts.

Three-dimensional models generally are made up of two parts. A stick figure component that represents the skeletal structure, and a volumetric (e.g. cylinders) representation for the surrounding flesh. Badler and O' Rourke [61], presented the earliest work in the field of motion recognition using a 3-D model. There are two basic approaches to the problem of obtaining a 3-D model, kinematic and dynamic models. Kinematic methods [22, 60, 61, 66, 70] , only consider the geometry of the objects, i.e. the position, orientation and deformation, without any consideration of the physical forces that may have caused the movement. Dynamic models [10, 51, 63, 81] take into account these forces and torques in order to produce realistic motion models.

2.3 Gesture and action recognition

Action recognition can be seen as a classification problem involving time-varying feature data [23]. The previous sections explained different approaches to obtaining the feature data. Recognition becomes the matching of an unknown test sequence with a known set of labelled actions.

Recognition of gestures or activities can be achieved by using state-space or template matching methods [1]. Template matching methods match the features extracted from an

image sequence to a pre-stored pattern. State space methods define each static posture as a state. An action sequence is a composition of a number of these states.

The work of Bobick [9] applied state-space approaches to the recognition of body actions, such as sitting, waving and crouching as well as hand gesture recognition. Campbell and Bobick [12] present a method for recognising classical ballet steps from 3-D range data. The movement is represented using space curves in subspaces of a "phase space". An action is defined as an ordered sequence of such states. Bregler [10] implemented a method using statistical models at various levels of complexity. Motion is recognised using a Hidden Markov Model (HMM), a probabilistic technique for the study of discrete time series. The technique is demonstrated by recognising gait categories in cluttered environments. A method for the recovery of the temporal structure of natural gesture is presented by Wilson et. al. [78] their work recognises natural gestural phases such as the rest transitional and stroke states that occur during conversation situations. A Markov Model where the states are not hidden is used for recognition.

Template matching approaches try to match features extracted from a sequence to a known pattern or template. Polana and Nelson [68] detect periodic motion such as a person walking that is recognised using low level, non-parametric spatio- temporal templates. The system evaluates an image sequence at pixel level to determine whether any of a known set of periodic activities is present. Bobick and Davis [8] use a two component temporal template, a binary motion-energy image (MEI) and a motion-history image (MHI). The view specific templates are matched against stored models of known movements to recognise the action. Masoud and Papanikkolopoulos [49] use an Infinite Impulse Response (IIR) filter to capture motion information directly from a number of sequences. These features are then mapped to a manifold in eigenspace and recognition is performed by calculating the distances to reference manifolds.

This concludes the review of influential work in human pose and action recognition. The research in this field all have one thing in common; they consider the space occupied by the person to recognise the pose or action. The following section contains a discussion on negative space, image regions surrounding a person, that forms the focus of this dissertation.

2.4 Negative Space: An unexplored research area.

This dissertation presents the notion that actions can be recognised by focusing solely on the negative space. Negative space is a fundamental concept in art theory as discussed in the next section.

2.4.1 Negative space

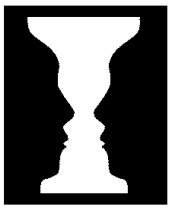

Figure 2.1: The Gestalt Chalice [14].

An image can be divided into two different regions the positive and negative space, sometimes referred to as figure and ground [4]. Figure 2.1 illustrates the interplay between these regions. The image can be seen as either a chalice, or two faces in profile, depending on whether the white or black areas are perceived as the object of interest.

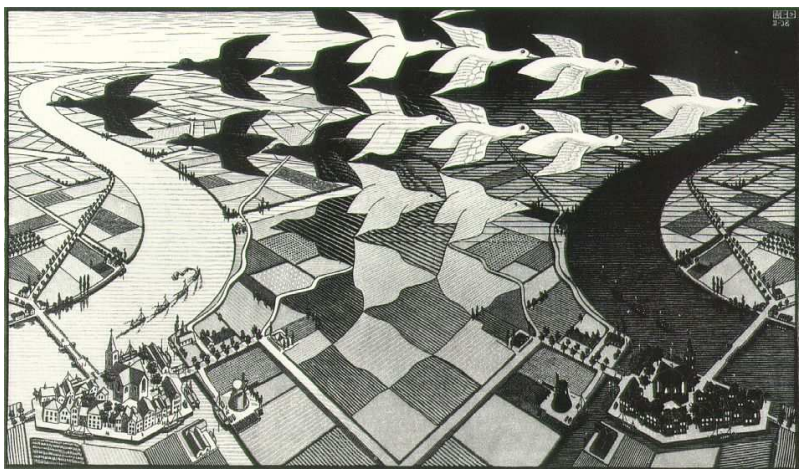

Figure 2.2: *Day and Night.* M.C. Escher 1938. Woodcut in black and grey [21].

Figure 2.2 is a woodcut by Esher, the middle of the image divides the space between two competing visual realities: whether the white or the black bird comes to the fore or

recedes. In the right hand side of the image the white bird is the positive space, the foreground, and the dark background makes up the ground, or negative space. The converse is true for the left hand side of the image.

Figure 2.3: *The Gray Tree.* P. Mondrian 1911. Oil on canvas [56].

The work by Mondrian, The Gray Tree, illustrated in figure 2.3, shows how the spaces between the branches are given more importance than the tree itself. Traditionally the subject of a painting forms the foreground, here we see the inverse, where the negative space becomes the focus of attention.

The work presented here relies almost exclusively on the background or negative space areas, instead of the region formed by the foreground, a human performing different actions. Analysis of these regions presents a novel method for the automatic recognition of human actions [38, 39].

2.4.2 Negative Space and the human figure

Figure 2.4: Segmented silhouette figures.

Figure 2.4 contains a number of silhouette images obtained from a person performing a waving action. In the waving sequence the actual person is the positive space, the object of interest as far as a human observer is concerned. The negative spaces are those white regions that surround the individual. The images are devoid of any colour texture or 3-dimensional shape information, but are merely a collection of pixels assigned either one or zero, yet the action portrayed can easily be recognised.

Figure 2.5: Images highlighting key negative spaces.

After studying a number of binary images from the waving sequence, it can be seen that the pattern formed by the negative space is more descriptive of the motion than the pattern formed by the positive space. Consider only the positive spaces: the person's torso, legs and head remain in approximately the same position while only the position of the arms change. There is very little change in the general shape of each arm and the pose information is almost exclusively contained in the angle that the arms make with the body. In order to recognise these poses one would have the difficult task of first extracting the positions of the arms and then trying to reconcile them with probable poses and actions. This would require extensive prior knowledge as well as methods such as model fitting or template matching.

Consider, however, the negative spaces surrounding the person; these areas drastically change shape and position, making it more characteristic of the pose. The images in Figure 2.5 highlight the changes in a number of these regions. The assumption, that useful information on a particular pose is contained in the negative space areas, resulted in the exploration of negative space methods for human pose and action recognition that are presented in this thesis.

A large part of the work was dedicated to finding a simple process to transform this visual solution into a measurable entity. A method had to be devised that would mathematically capture what seems so clear visually. Several approaches were considered to this problem. For example, the negative space regions could be characterised by contour descriptors or it could be simplified as represented by geometric centres of mass. The method which proved ultimately most efficient i.e. the simplest characterisation retaining the most information, was to consider the negative space in terms of the mathematical areas it creates around the human silhouette.

It should be noted, however, that the separation of the positive to negative information of the human silhouette involves a segmentation process and segmentation itself is an unresolved problem. Therefore, this study has circumvented those issues by focusing on action sequences captured in a controlled environment.

Chapter 3

The image sequence data

The development of a human action recognition system requires a database of people performing different actions. To investigate the negative spaces we have to extract a silhouette image of the individuals in the image sequences.

To obtain the human silhouettes we need to separate the image pixels belonging to the human from the image background, a process referred to as segmentation. It is possible to simplify the segmentation process by using Chroma-keying, or blue screen, techniques [73].

The data for this dissertation was collected at two different locations. The Chroma-key room at the University of Cape Town and in a professional Chroma-key film studio.

The two locations required different data capturing methods. The film studio footage was shot on digital tape in a single day. The UCT studio has been available since its construction for on-line sequence capturing and segmentation.

The bulk of the experiments were performed in the UCT studio room built on campus. The space allows for only a single camera view. Later stages of the work required data from multiple angles and was filmed in a professional film stage built according to the requirements of this dissertation.

3.1 Dataset 1: UCT Chroma-key studio

The UCT dataset contains images acquired in a custom built chroma-key studio. Figure 3.1 shows a schematic representation of the room. The room comprises a Chroma-key stage area, a computer station, a camera and a lighting grid. The walls and floor within camera view are covered in blue Chroma-key fabric. The computer station houses a PIV, 700MHz PC, a frame grabber and 30 Gig storage space for the captured sequences.

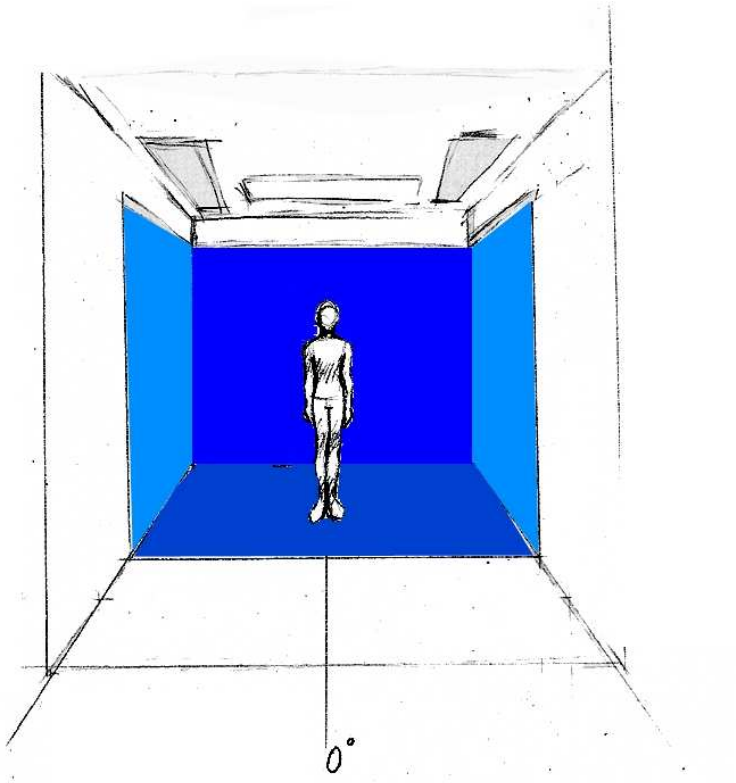

Figure 3.1: The layout of the UCT Chroma-key studio. This space can accommodate only a single front view camera.

The most critical part of any blue room is a uniform coloured background. This relies not only only on the appropriate choice of colour but also on the lighting. The lights must be placed in appropriate locations to minimise the amount of shadows cast by the filmed subject/s and to evenly light the background walls. The subject and the background must be lit separately. A special “lighting grid “ was constructed on the ceiling of the room to provide lighting. The lights can be moved to various locations on the grid to enable

different lighting combinations. A number of fluorescent lights called “wall washers” are used to evenly illuminate the walls. Diffused tungsten lights are used to illuminate the subject. Figure 3.2 shows an empty frame while Figure 3.4 shows a number of images of people in front of the blue Chroma-key background.

Figure 3.2: Background UCT Blue Room.

The UCT Chroma-key room has proved to be a invaluable experimental tool. Digital image sequences can easily be captured and segmented to quickly test hypotheses without the need to set up filming equipment and implement complicated segmentation algorithms.

Figure 3.3: The training set

The action recognition algorithm described in this dissertation was developed using image sequences from the Chroma-key studio. Two different women performed 15 minutes of random free-form motions, these varied actions comprised the training sequence. A number of frames from this sequence can be seen in Figure 3.3 .Figure 3.4 shows a few examples of people performing actions from the testing set.

Figure 3.4: Sequence examples.

3.2 Dataset 2: Film Studio Chroma-keying

The work soon outgrew the limitations of the UCT studio and a new source of sequence data had to be found. The UCT studio has the disadvantage that actions can only be viewed from one camera angle. A larger studio had to be sourced to investigate changes in the negative space when an action is viewed from different angles. This section discusses the construction of the multiple camera stage, organising the actors and their actions as well as managing the data.

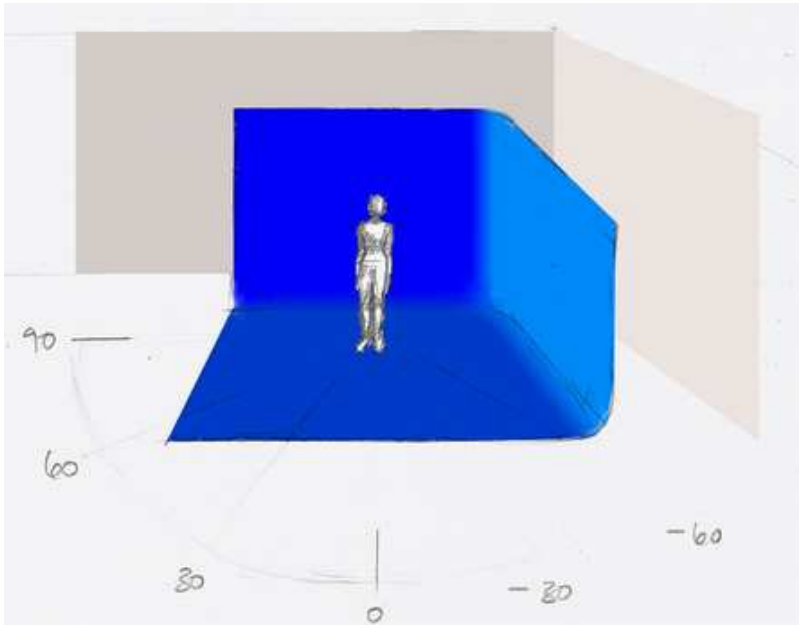

Figure 3.5: The multi-view Chroma-key stage.

3.2.1 The Chroma-key stage

The set had to be constructed out of wood and painted 4 layers of Chroma-key blue requiring 4 days of preparation. The floor of the set was 6m by 4m, and the side walls were 3m high. The floor and sides as well as the joining of the two side walls were formed using infinity curves, which are rounded corners that allow for an even colour transition between floor and sides to simplify later segmentation. The lighting was set up before hand and included some unavoidable natural lighting cast on the scene through the windows.

The data was recorded on 6 different cameras. The cameras filmed the actions continuously and simultaneously on mini-DV tapes. Each of these cameras have different gains as well as noise levels. These differences in camera characteristics and lighting can be seen in the variations between the different blue backgrounds of the images contained in figure 3.6.

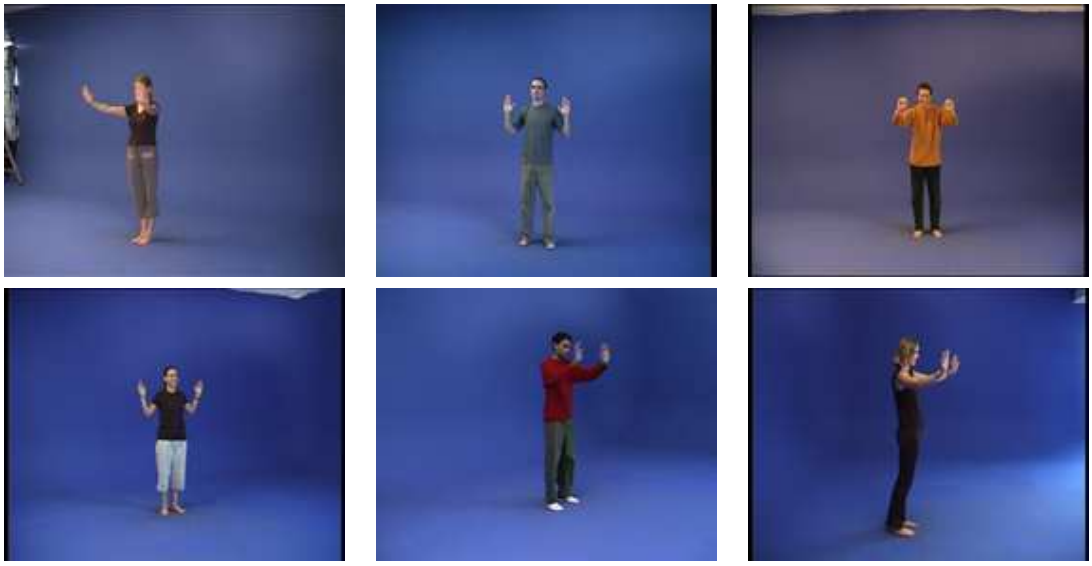

Figure 3.6: Video stills. The first two rows show images from the different camera angles. The camera angles were at 0, +-30, +-60 and 90 degrees respectively.

The six cameras were stationary and arranged in a half circle shown in figure 3.5. To the left there were 3 cameras at +90, +60 and +30 degrees. The centre camera is marked 0 degrees and the cameras to the right -30 and -60. There is no -90 camera angle, as this camera view is covered by the +90 degree camera. In the silhouettes, these views provide the same information.

3.2.2 The Actions

The actions were similar to those used in the work by Bobick and Davis [8], but were extended to include actions such as lying down, waving, beckoning etc. The data set deviates from this set as we have not used trained professionals. The benefit is that the system could be more robust to untrained data as the training set provides a far greater deal of variation.

There were a total of 7 people performing the actions, three women and four men. The variation in the manner that different individuals perform the actions could form a more representative pose set. However, it does complicate matters at the recognition stage. If the individuals had more time to rehearse the moves they would appear more similar. Thus, some of the same actions performed by different people appear to the camera significantly different, and it will be difficult at the recognition stage to decide if a false recognition should be ascribed to these variations.

There was no restrictions placed on the type of clothing that the participants could wear, except that it should not be blue and no dresses were allowed.

It was not possible to predict beforehand the amount of actions that could fit into a one hour mini-DV tape. This consideration formed the basis for the action set approach. The actions are performed one at a time by all the people in the set. Each of these individuals in turn repeat the actions eight times. For example person one will enter the space and repeat action one, eight times. He or she then leaves the camera field of view and person two enters to repeat the action.

The action approach is beneficial from a data structuring perspective, since a single tape would carry all the data necessary to train and recognise several actions. It also allows the designer to decide how much data to allocate toward training and testing.

3.2.3 Timing and Temporal Labelling

At the planning stages of the data set it was clear that the amount of information would be far too vast to label manually. However, a labelled set is crucial to automatically asses how well the system recognises the desired actions or poses. It is too complicated to label the data at pose level, but action labelling is made easier by how they are organised on the tapes.

Grouping the data according to actions and the different individuals performing them simplifies not only data management on the tapes but also the file management problem on the computer. This is an important consideration as the data need to be structured to make it more manageable. The digitised sequence frames have to be placed in different

folders small enough so that the content can be viewed with explorer as well as various image viewers.

Programs were written for 2 laptops, which were used to temporally segment the data. The operators have to enter when a new action starts, a different person enters the room, and the start and end of each of the individual action repetitions. This timing framework was saved in a matrix with entries corresponding to these different key times. This timing matrix needs to be offset for different actions as there is often a slight delay relating to human reaction times. The timing is relevant for the tapes from all six of the different camera views obtained during the same filming session.

3.2.4 Processing the data

The video footage was recorded on mini DV Tapes, each of these contained an hour of video, giving 24 hours in total. This data has to be transferred from the tape to an image sequence that can be used for MATLAB processing.

The mini-DV cassettes are smaller versions of DV cassettes that apart from physical size and shorter running time retain all the properties of the aforementioned medium. DV compresses the signal to a fixed compression of 5:1 which is a moderately lossy compression. This equates to about 3.6 MB/sec or approximately 100 Kb per frame. The compression is a variant of the Discrete Cosine Transformation, the basis of JPEG and MPEG.

The mini-DV tapes are read into the computer via Firewire from a SONY DCR-PC5E digital video recorder. Adobe Premier 6.0 has proved very useful in capturing the data from the tapes. It has a DV device control feature that allows for batch-capture of video. A timing matrix described in section 3.2.3 containing the start and end times of each individual action repetition is imported into a batch file. Premier has control of the DV device from which it automatically seeks, records and saves the segments.

3.3 Segmenting human silhouettes from the Chroma-key background

The previous sections discussed how the image sequences were obtained. This footage still contains the blue Chroma-key background. The following section explains how the Chroma-key technique is used to segment the different datasets.

The term Chroma-key or bluescreen refers to the practise in the film industry of filming an object in front of an even coloured background that can easily be "keyed" out, leaving only the silhouette image, or mask, of the foreground object [73]. Chroma-key creates keys on just one colour channel, the blue in this case. Different colours such as red and green can be used, but blue is mostly favoured when working with people as it is complementary to flesh tones. In digital applications the background can be segmented using simple thresholding. A limitation to Chroma-key techniques is that the foreground object should not contain any of the background colour. Blue spill refers to the reflected light from the background onto the subject giving them an unnatural tinge. to obtain a good key with low noise, the background should be illuminated both brightly and evenly.

3.3.1 UCT Blue Room

The blue pixels, corresponding to the Chroma-key background are identified and set to zero. By masking out these areas only the foreground pixels retain a value above zero. By setting these foreground pixels to one we have created the binary silhouette needed to investigate the negative space.

3.3.2 Film Studio Dataset

The film studio set did not have controlled lighting. The algorithm used in the UCT studio is adequate but have to be modified to accommodate the different camera angles. The modification is necessary as the lighting conditions as well as camera characteristics are not identical for the different views. Figure 3.6 shows some of this variation corresponding to images taken during from the same action.

The lighting was not controlled and included natural lighting. As the four hours of filming progressed there is a notable change in background characteristics due to the moving sun. Small adjustments in the algorithm can compensate for this variation. The algorithm works on the basic principle that there is a unique ratio of red to blue , (or blue to green depending on the image), for a given chroma key background. Shadows for the most part only cause the intensity of the colour to decrease, the ratio remains the same.

After segmentation the binary image is labelled and the largest connected region near the centre of the image is extracted. It is not possible to use a filling procedure to include small isolated mislabelled regions within the body of the person, as it is virtually impossible to distinguish these areas automatically from legitimate "holes" formed by the spaces between the limbs and the body.

Chapter 4

Negative Space Analysis: preprocessing and feature extraction

This chapter describes how the human silhouettes obtained in Chapter 3 are processed to eventually recognise the poses. Areas of the negative space, surrounding the silhouette figure, form patterns when a person moves that are more descriptive of the action than the pattern formed by the silhouette itself.

Figure 4.1: Segmented silhouette figures.

Figure 4.2: Images highlighting key negative spaces.

4.1 Extracting the negative space areas

To highlight the negative spaces surrounding the silhouettes a preprocessing method that colour codes the negative space was developed. This uncomplicated preprocessing method simplifies the feature extraction and pose recognition stages. By using a scanning process described in this section, it is possible to isolate these regions. This preprocessing step significantly enhances the dissimilarity of distinct poses. A rectangular bounding box is constructed around the extremities of the silhouette image, and only negative space regions within this bounded area are considered. The rectangular bounding box, instead of another geometric shape, was chosen due to its simplicity to implement and physiological experiments that show that the ratio of a persons horizontally outstretched arms to their vertical height shows little variation for different individuals.

Within the bounding box the negative space areas are more numerous, and undergo more radical changes in shape than the single area occupied by the figure of the person performing the action. A method was devised that could capture these changes in the different negative space regions. This preprocessing method identifies distinct negative spaces by assigning to them a unique colour code. This colour coding of the negative space is an image preprocessing tool that characterises each of the distinct regions between the body parts and the bounding box.

The negative space preprocessing method requires a binary silhouette image as input where the pixels occupied by the person are set to zero and the surrounding spaces set to one. A horizontal scanning process is used as a first stage method to extract the negative spaces contained within the bounding box. This horizontal scanning or filling process can be described as follows: the image is scanned from left to right, any pixel with a value of 1 is set to 0 until the first zero valued pixel (the person) is detected. This is done for each row in the image. In a similar manner the image is also scanned from the right to the left, isolating negative space areas formed on the right hand side of the body. Figure 4.3 shows images resulting from these two procedures that were binary anded together for illustration. Binary addition results in a combined image where the negative space areas formed between the arms and the body as well the legs are clearly visible.

Figure 4.3: The images illustrate the combined results of the horizontal scanning process.

Images formed by the horizontal scanning method clearly show how distinct pose patterns are formed by the different poses in the action. The visual power of the horizontal scanning provides the motivation for a similar vertical scanning process. Two vertical scanning methods are added, one image resulting from scanning the silhouette from the top to the bottom, and another by scanning from the bottom up.

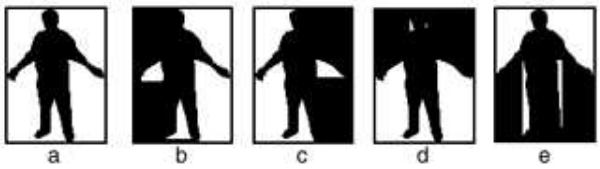

Figure 4.4: Individual images that result from the horizontal (b,c) and vertical scanning (d,e) processes.

4.2 Visualisation of the preprocessed silhouettes: Colour coded images

We now have four different images generated by a single binary silhouette image: two horizontal and two vertically scanned images. Instead of having the original single image, we now have four different images. These images have to be incorporated into a single image to visualise their combined negative space information.

The four different sets of data resultant from the scanning processes can be combined by assigning them to the three separate colour fields, the red green and blue fields of an RGB image. As there are four images, two of the scans will share a colour field. It makes no difference to what colour field the different scans are assigned as the colours are mainly for visualisation purposes. The vertical scans shown in Figure 4.4 (d) and (e), are multiplied by 255 and assigned to the green and blue fields of the RGB image.

Since there are two horizontal scanned images, Figure 4.4 (b) and (c), this requires the allocation of both binary images to a single colour field. The left scan is multiplied by 185 and the right scan by 70. These scans are added together to form the red field. The different multiplication factors are chosen arbitrarily to produce the most visually distinct result.

Figure 4.5: The Colour coded image that is formed by combing the different scans in Figure 4.4 into a single RGB colour image.

The result of this operation is that the different scans are unified into a single, coloured image, termed for obvious reasons the colour coded negative space representation. Images resulting from this procedure now not only contain the original negative space areas, but interesting new patterns arise where the scans overlap to form new coloured regions as can be seen in Figure 4.7. A total of 10 different colours are in fact possible.

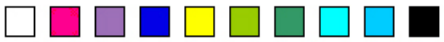

Figure 4.6: Possible negative space colours

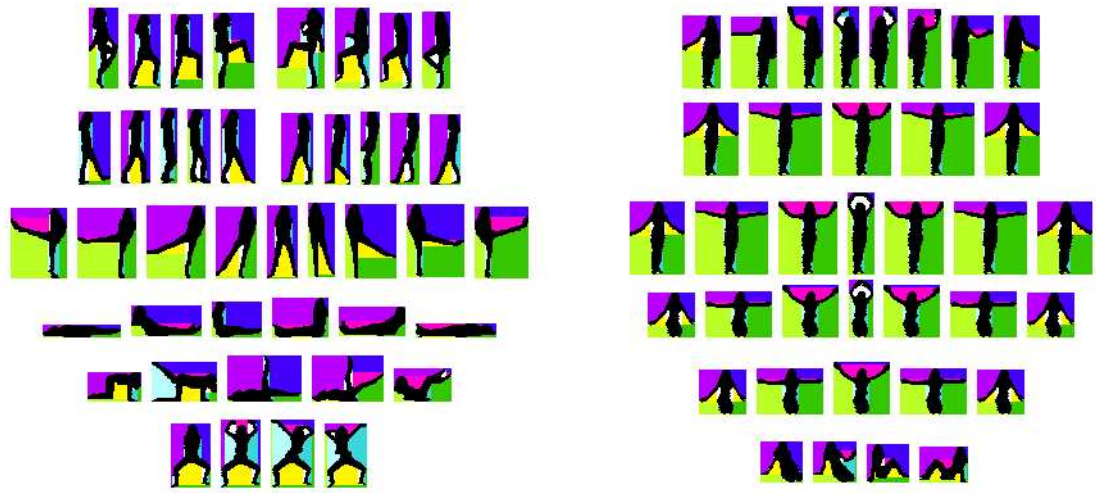

Figure 4.7: Colour-coded negative space images showing the characteristic patterns that arise for the distinct poses.

A further processing step replaces white areas bordering on coloured regions to increase the amount of characteristic colour. The white areas are replaced by the colours bordering the sides of these regions. Figure 4.8 show colour coded images before and after replacing the white regions. This operation leaves only the white areas that are surrounded by black untouched, Figure 4.9 shows a number of these examples.

Figure 4.8: The white regions are filled with the colours of the area bordering its sides, this increases the space occupied by the colours and reserves white only for those areas completely bordered by black.

Figure 4.9: Images where the white regions are left unchanged.

Colour-coded images as depicted Figure 4.7 show that this representation does indeed capture variations in different poses. For example key variations between a wave and a crouch or a leg kick can be seen in the dominant colours. If a family of waves is considered, the coloured regions alone can classify the differentiating poses.

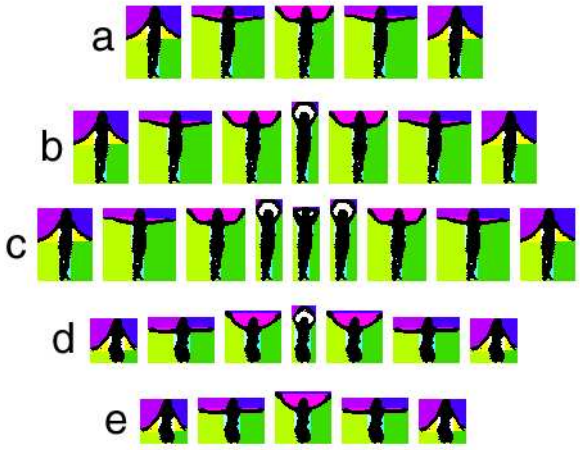

Figure 4.10: Colour coded negative space images representative of the wave family

The wave sequence in Figure 4.10 a) show that a characteristic red region is formed between the two uplifted arms. This colour code is very distinct from that of Figure 4.10 b) when the arms move over the head and a white region gets formed that is not present in the previous wave sequence. The same characteristic pattern emerges when a person sits down and waves, however, the bounding box proportion as well as the amounts of green negative spaces now present distinguish this pose from the previous similar standing wave pattern. Thus, different characteristic patterns formed by the coloured regions and the bounding box distinguish the different poses.

4.2.1 The colour coded image databases

The colour coded images in Figure 4.19, 4.17 and 4.15 are colour coded negative space images from Database 1 described in Chapter 3. These images show how similar the coloured regions are when people perform the same actions. Unfortunately the full strength of this method cannot be appreciated in static images. The Appendix CD contains a large number of animations made from the colour coded images of both the dataset

sequences described in chapter 3. These animations should be viewed before reading the following sections.

4.3 Extracting features from the negative space images

The previous sections illustrate how the negative space preprocessing method has highlighted the differences between poses. The individuals in the images have very different clothing and appearance, yet the colour coded method visually captures the similarities between the poses. The colour-coded negative space representation is mainly used for visualisation purposes; actual feature extraction involves assigning a unique number to each of the possible colours. The next step is to extract features from this representation that can be used to classify the pose.

4.3.1 Approaches to feature extraction

Features have to be extracted from the colour-coded negative space images to classify them. Many approaches were considered, from template matching to neural network methods. However, each of these approaches needed some form of labelled data to proceed. Classifying the data in order to obtain templates or training data is difficult, not only due to the large number of images from different people performing the actions, but also because of the difficulty in choosing representative poses. There is no clear way to label the poses, any attempt to do so would impose a structure on the data that may not exist. The problems described imply that for this thesis a method which relies in any way on the pre-labelling of pose images would not be considered.

Different poses are distinct due to the shape and positions of the coloured regions, as well as the proportions of the bounding box. Initially, it was thought that shape descriptors needed to be extracted as features to capture the essence of the images. This approach would have involved complicated methods such as b-splines or snakes, as well as additional features describing the spatial location of these regions.

The most attractive feature of the colour coded representation is its simplicity. It is simple to implement and to study.

4.3.2 Feature extraction from the colour coded images

Figure 4.11: Spaces from which different colours can arise.

A study of various colour-coded images led to the conclusion that pose information is contained in the size of the coloured areas and whether they are present or not, rather than in the actual shape of these regions. There is no need to identify the actual spatial location of these coloured areas as the colours themselves imply a spatial location; consider for example the light green areas at the side of the body in Figure 4.11 illustrating when both the arms are extended at the sides, and when only one arm is extended. The light green colour always originates from the bottom left hand side of the image, therefore the percentage of colour indicates to what extent the body 'opens up' toward this side. It is tempting to want to include at least a description of whether this area is tall and thin or perhaps more squarely shaped, but some thought should reveal that these additional features become unnecessary when considered in context with the rest of the image. All the coloured areas are interdependent. No coloured area can simply become larger without resulting in an additional change in the bounding box proportions or percentages of the other coloured areas. Thus, spatial information about one region is also implied by the percentages of the other colours present. This strong relationship between the coloured regions suggested that the percentage of each colour present could be extracted as features for recognition.

An interesting relationship arises between the percentages of the body present vs. that of the coloured areas. When a person extends their arms to the sides as the percentage of negative space surrounding them is at a maximum, but for all other poses the bounding box size will decrease and the ratio of coloured regions to the person will decrease. Thus a person sitting with their legs crossed might have the same bounding box ratio as a person standing with their arms outstretched, however, the ratio of negative space to body area will be greater in the latter case. This relationship makes it possible to recognise different

poses without having to know the relative sizes of the bounding boxes within the physical image space. No information other than that which is available in the image itself is used. This meant that no reference height or 3-dimensional location of the individual was considered.

Images are distinct not only because of the pattern formed by the coloured regions but also due to the dimensions of the bounding box. When a person is standing the bounding box is elongated in a vertical sense. A person lying down would have a bounding box orientated horizontally. This ratio of height to width is expressed as the angle that is formed between the diagonal and the base of the bounding box. The angle $\theta = tan^{-1}\left(\frac{h}{w}\right)$ as shown in Figure 4.12 rather than the direct ratio $\frac{h}{w}$ is used as it provides a linear description of the bounding box proportions. The graph in Figure 4.13 illustrates that the angle θ provides a linear bounding box description while the ratio of height to width does not.

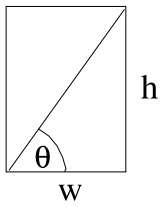

Figure 4.12: The angle shown in the diagram is used as the bounding box descriptor.

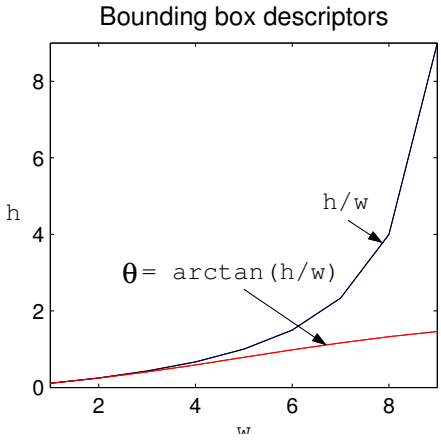

Figure 4.13: The plots show why theta is preferred for its linearity to the ratio usually used as bounding box descriptor.

The feature vector now consists of a total of 11 entries. These entries are the 10 colour

features, including black, as well as the bounding box shape descriptor . Figure 4.14 shows traces from a wave and kicking sequence. These plots illustrate how the colour percentages change as the person moves between the different poses that makes up an action. These traces provide motivation for the strength of both the colour-coded negative space representation as well as the feature extraction method.

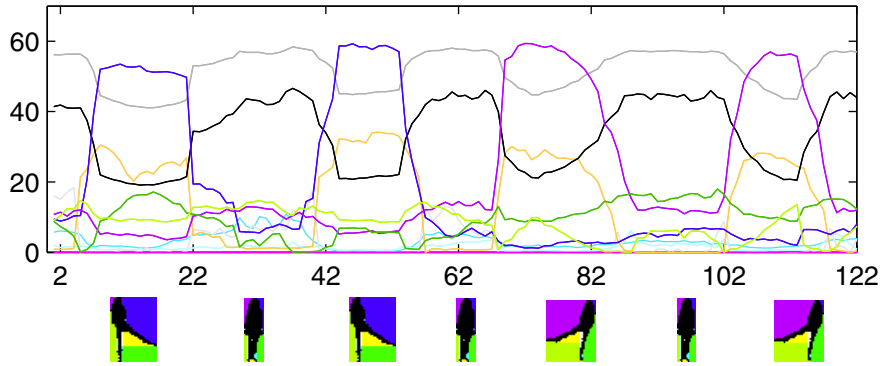

Figure 4.14: Selected features from a wave and kicking sequence. The colours in the graphs correspond to those found in the images.

4.4 Investigating the feature plots

The section contains a discussion of the feature plots from Datasets 1 and 2. Each image in the datasets are represented by a vector consisting of 11 features extracted from the negative space representation. The colour coded image representations made it possible to visually assess the strengths of analysing the negative spaces for pose recognition, however these images are now described by only eleven numbers. The feature plots are used to evaluate whether the extracted features have the same ability to capture the pose transitions within the actions. It is impossible to visualise such a multi dimensional space. By plotting the different features as traces simultaneously on the same axes, however, we can obtain a sense of how these variables relate to one another.

Each of the datasets contain a number of people performing a series of repetitions of different actions. The plots group the data from the action sets together to compare the deviations within a specific action. The vertical axes denote the percentage colour present and the horizontal axes represent the number of frames in a particular sequence. The different

traces represent the percentage colours present in the colour coded negative space images, black represents the space occupied by the body of the person and grey the bounding box shape descriptor. A number of colours are weighted to ensure that their characteristic patterns are visible in the plots along side the more dominant colours. The weightings are used throughout and are as follows: lime, olive, violet, maroon, light blue and dark blue are multiplied by 1.5. Yellow and red are multiplied by 3.5, white is multiplied by 4. Theta and black are multiplied by 0.7. Traces from Dataset 1 in Figures 4.16, 4.18 and 4.20 show a single repetition for each person performing an action, whereas the traces from Dataset 2 in Figures 4.21 to 4.24 show seven different people performing a succession of repetitions of a specific action.

4.4.1 Front view data plots from Dataset 1

The images in Figures 4.15, 4.17 and 4.19 show some of the key poses that make up the actions in Database 1.

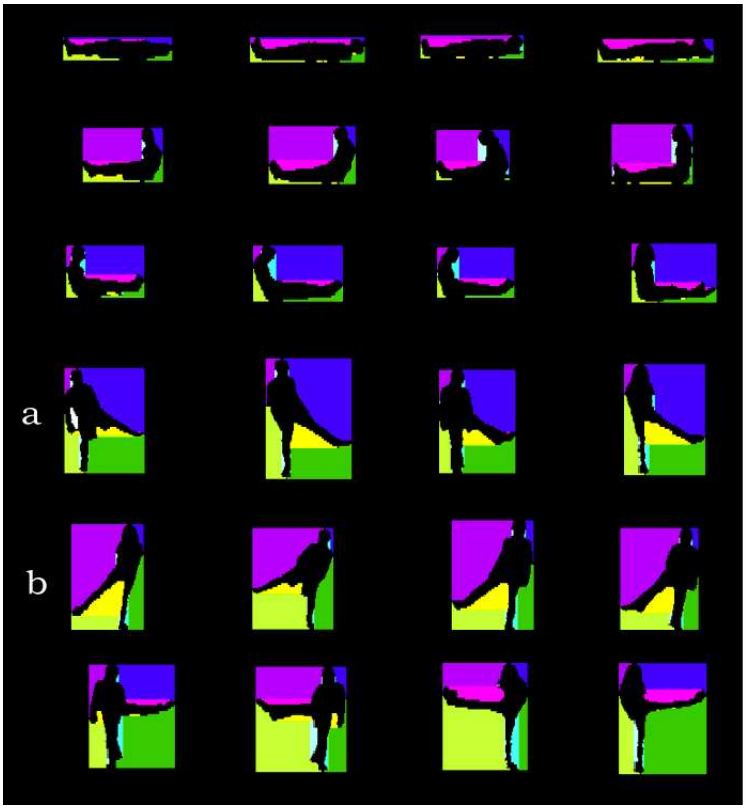

Figure 4.15: Key poses derived from actions in Database 1.

Figure 4.16 shows traces from Dataset 1 where the actors perform kicks to the side of their bodies, typical key frames for the different individuals are shown in Figure 4.15 a) and b). The traces from Action 9 in Dataset 1 are on the left hand side of Figure 4.16 and illustrate the increase in violet corresponding to that shown in the images of Figure 4.15 a). The Traces in Action 11 show the increase and decrease of maroon as the leg is lifted and dropped.

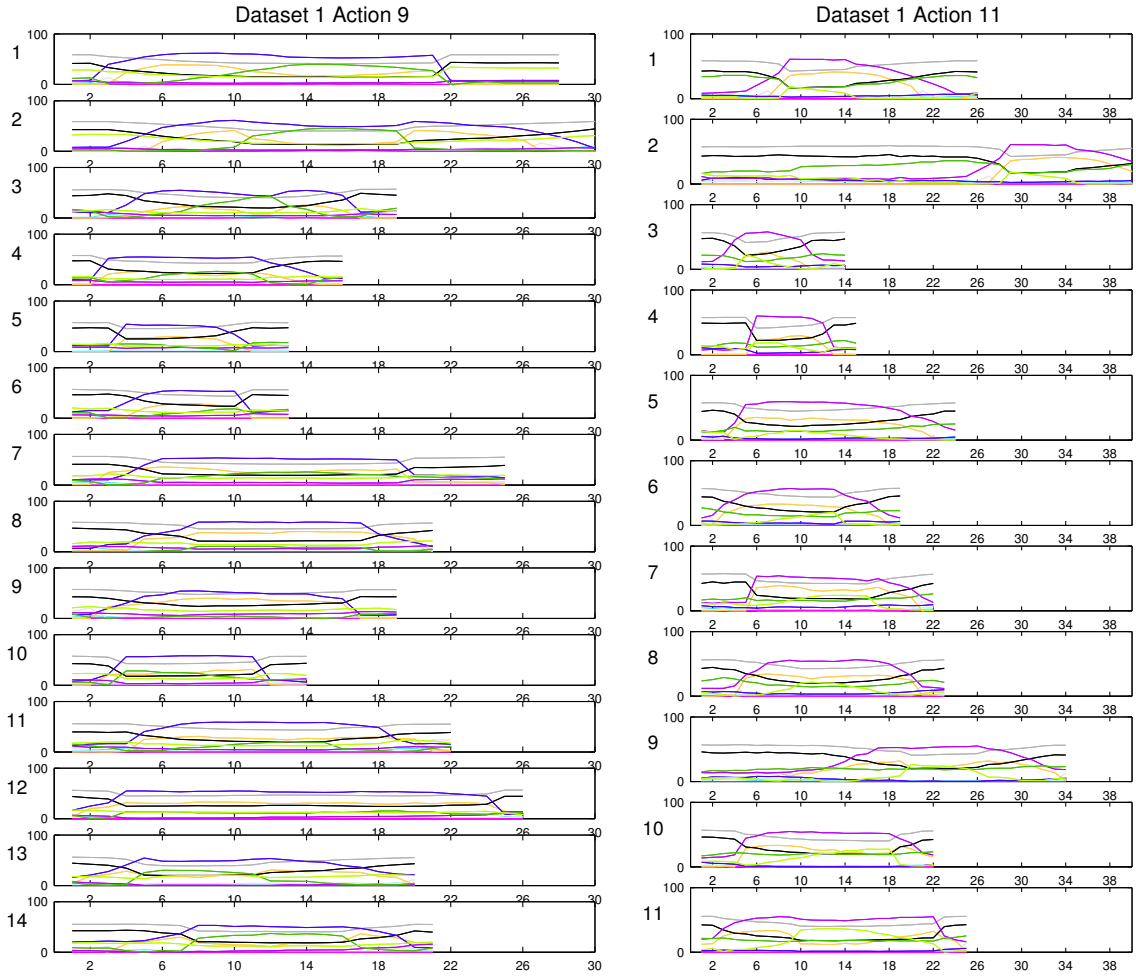

Figure 4.16: Traces from Dataset 1 showing right and left hand side kicks. They can be identified by the large amounts of yellow and either violet or maroon.

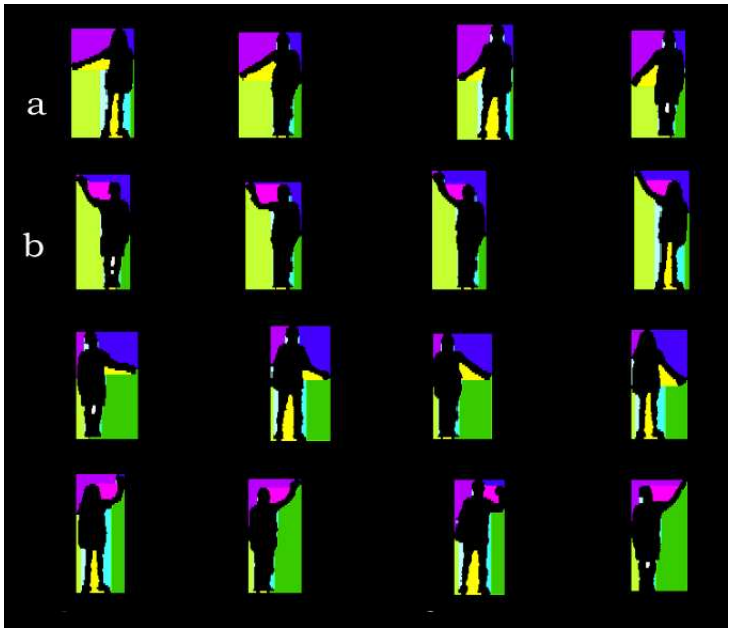

Figure 4.17: Key poses derived from actions in Database 1

Figure 4.18 shows traces from actions 3 and 4 where the right arm is lifted up to the side of the head in Action 3 and then dropped again to the side in Action 4. Action 3 corresponds to a person going through a stand to pose a) in Figure 4.17 and then to pose b). Action 4 contains the dropped arm sequence going from key poses b) then a) to a standing position.

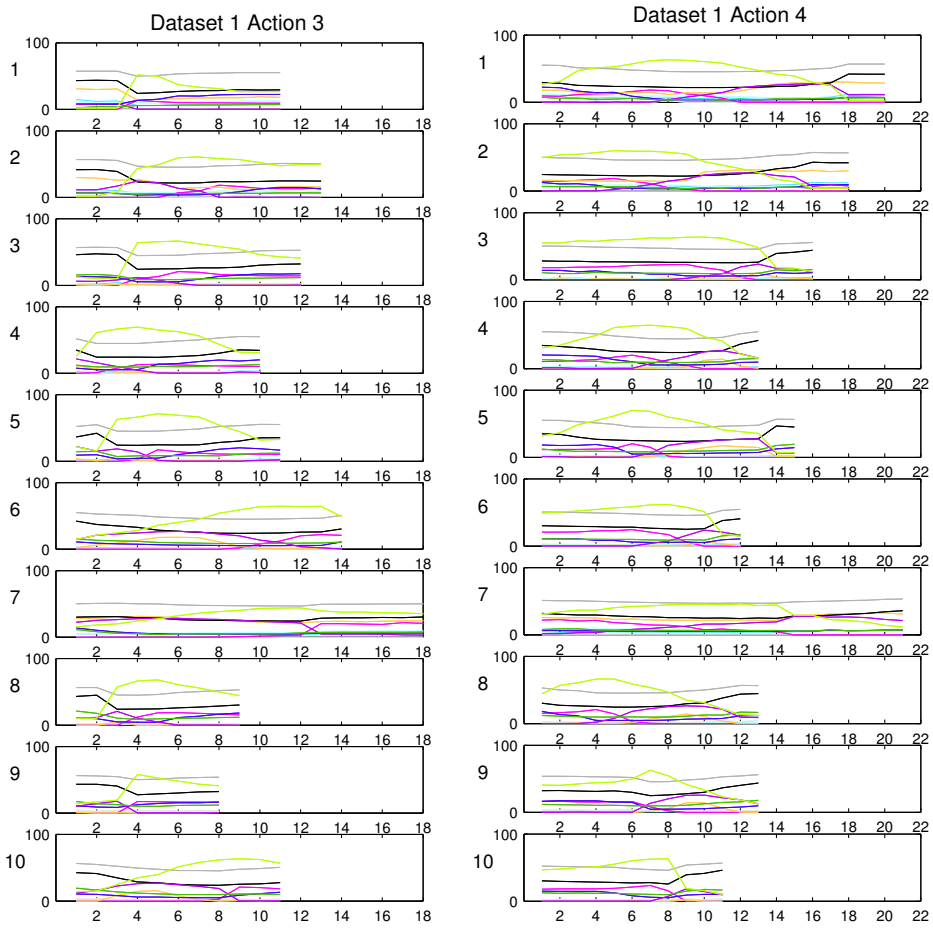

Figure 4.18: Traces showing variation within the same action group. The traces on the left shows the right arm being lifted, whereas in the right hand side plots the arm is lowered to the side.

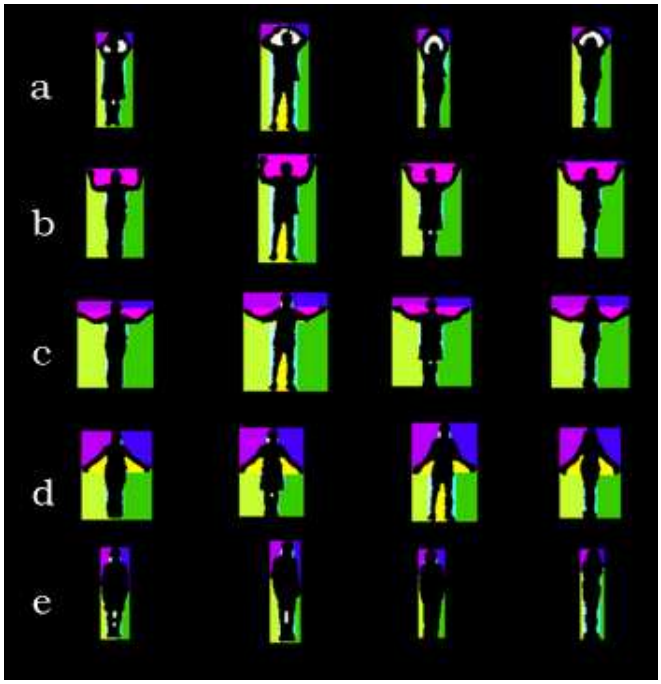

Figure 4.19: Poses of a wave action in Database 1

The traces from Action 13 in Figure 4.20 show a number of people performing a full wave that turns above the head. The key poses for this action are illustrated in Figure 4.19, the action moves from pose e) to d), c), b), a) and then back to e). The traces in Figure 4.20 show how differently individuals perform this simple action. The same colour relationships hold true for all, the actions start and end with a stand pose as key pose e) shows there is a large amount of black, the persons body in these images. This amount of black decreases as more negative space colours fill the box when the wave opens up. The percentages of lime and olive increase and decrease with the same amounts as this is a symmetric pose. The two red bumps that are visible in the traces come from regions shown in Figure 4.19 b), these bumps decrease when the arms are above the head and the key pose is similar to that in row a).

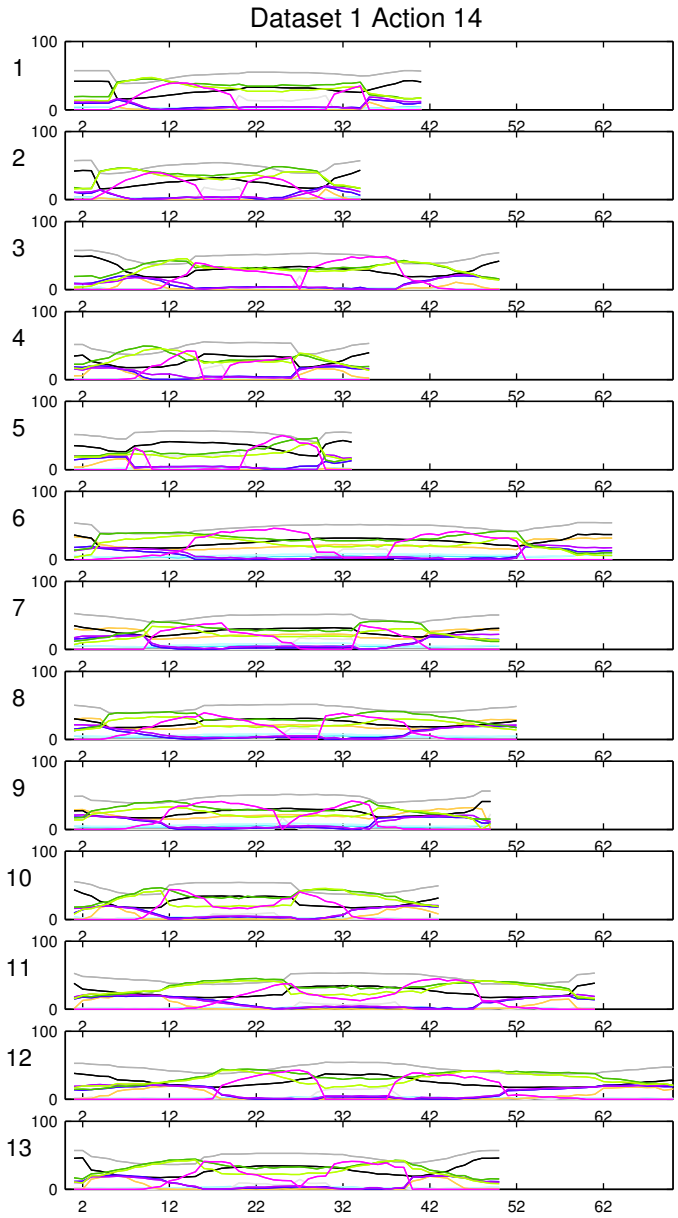

Figure 4.20: Traces from Dataset 1 showing a set of waves. The characteristic red bumps are formed when the arms are above the shoulders and decrease between the bumps when they are above the head.

The differences between the Action traces can be attributed to temporal variations as well as variability in the manner in which different people perform an action. Even if two actions were performed in exactly the same way they could appear dissimilar, due to noise introduced by the segmentation process or sampling differences resulting from the frame rate.

4.4.2 Front view data plots from Dataset 2

Traces depicted in this section are plots from sequences where the camera angle is parallel to the action plane. The following section contains feature plots of the same action taken simultaneously from multiple camera angles. The plots are too numerous to include all here, the rest can be found in Appendix A and B.

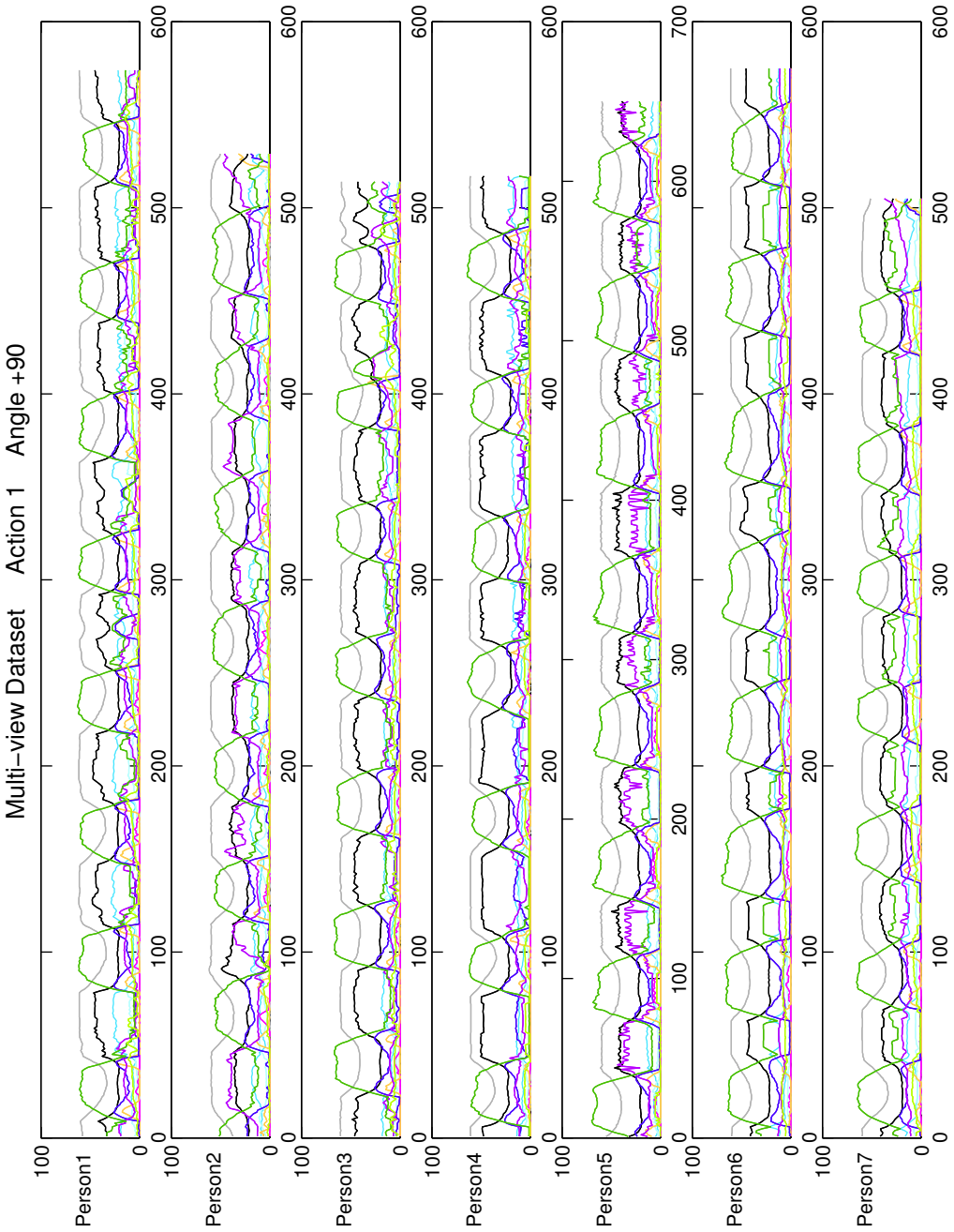

Figure 4.21: Feature traces from Dataset 2 showing distinct plots for 7 people performing a number of repetitions of Action1.

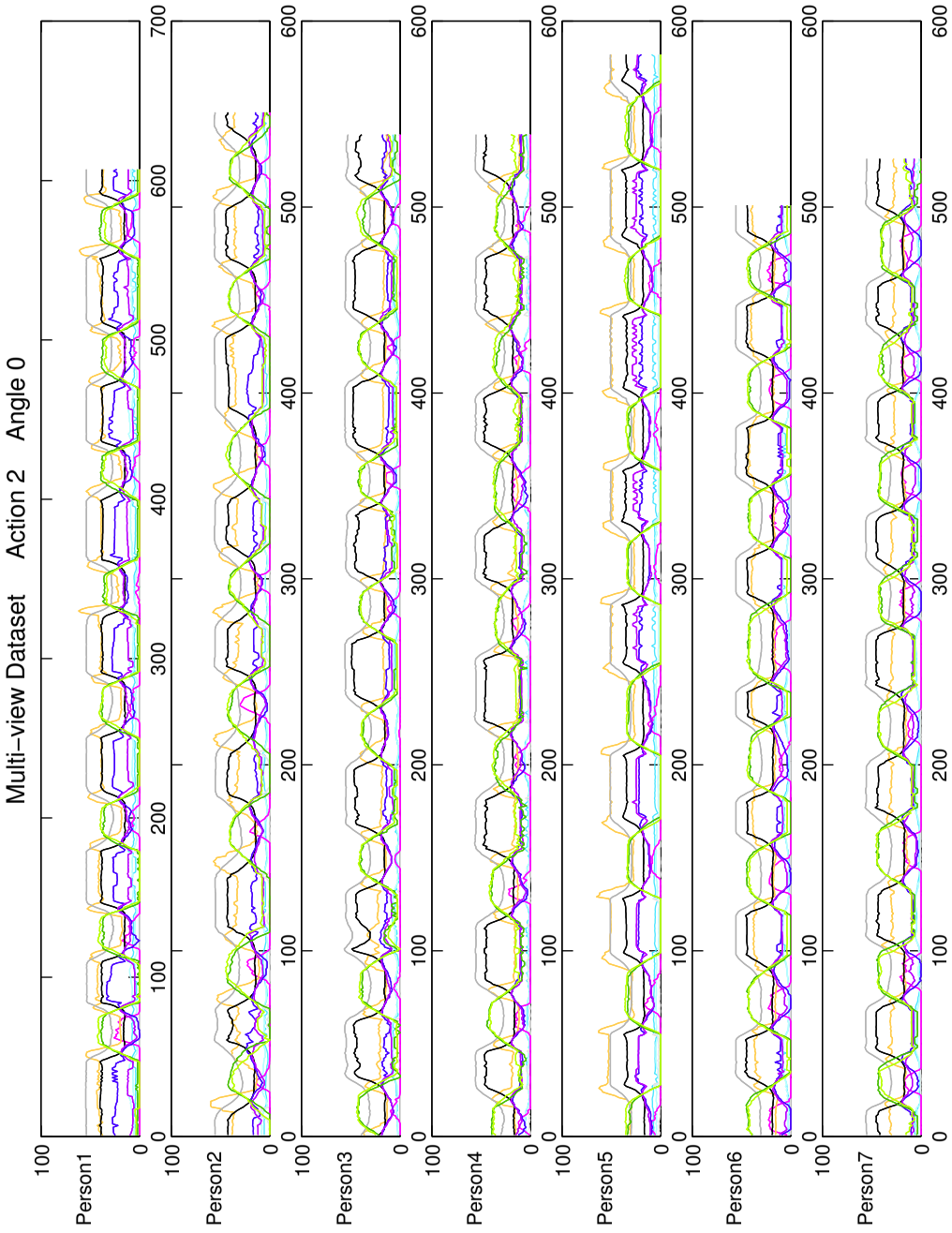

Figure 4.22: Feature traces from Dataset 2 showing distinct plots for 7 people performing a number of repetitions of Action2.

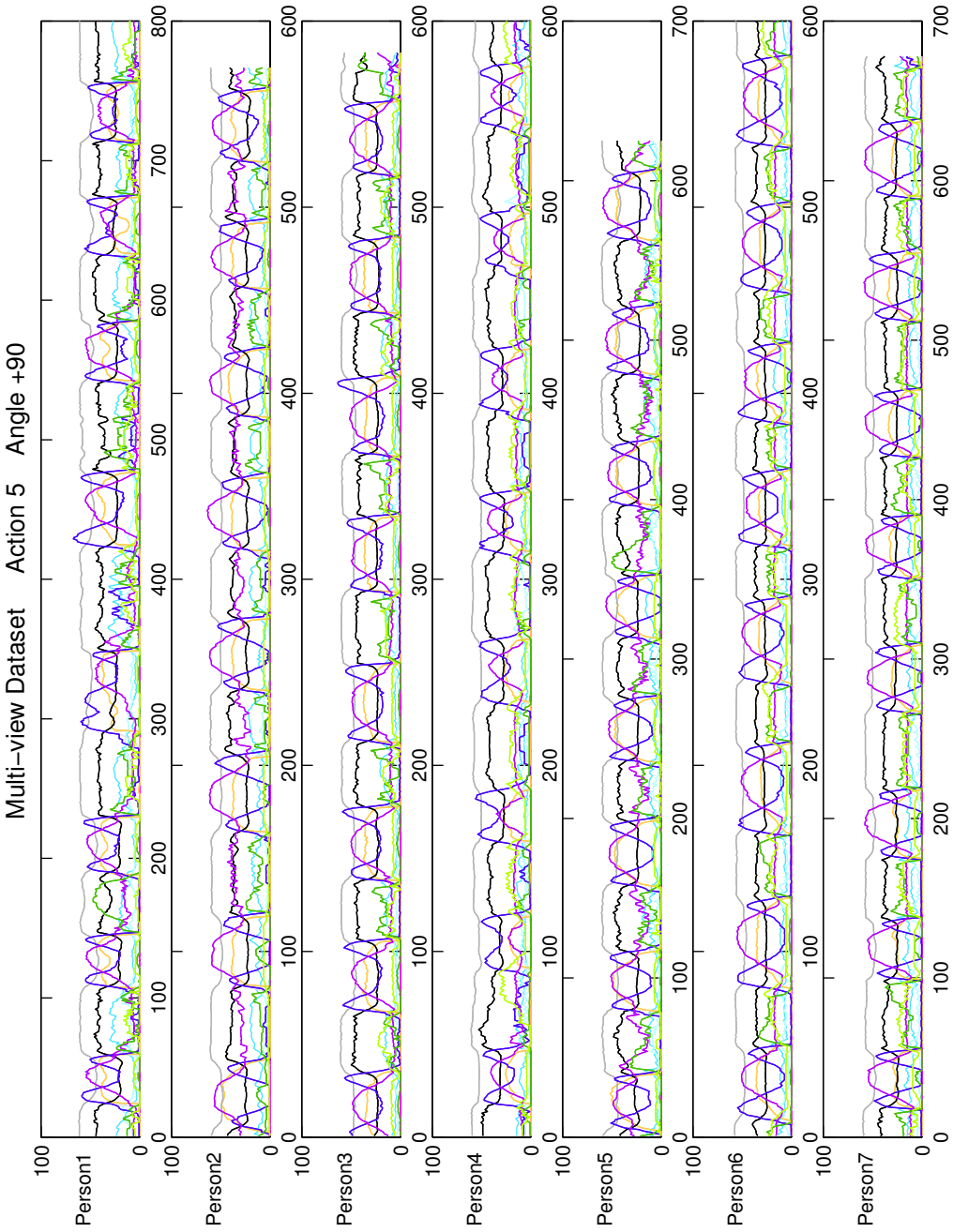

Figure 4.23: Feature traces from Dataset 2 showing distinct plots for 7 people performing a number of repetitions of Action5.

4.4.3 Multi-view plots from Dataset 2

The plots in Figures 4.24 to 4.29 show the how the feature traces for one action group varies when viewed from the six different angles.

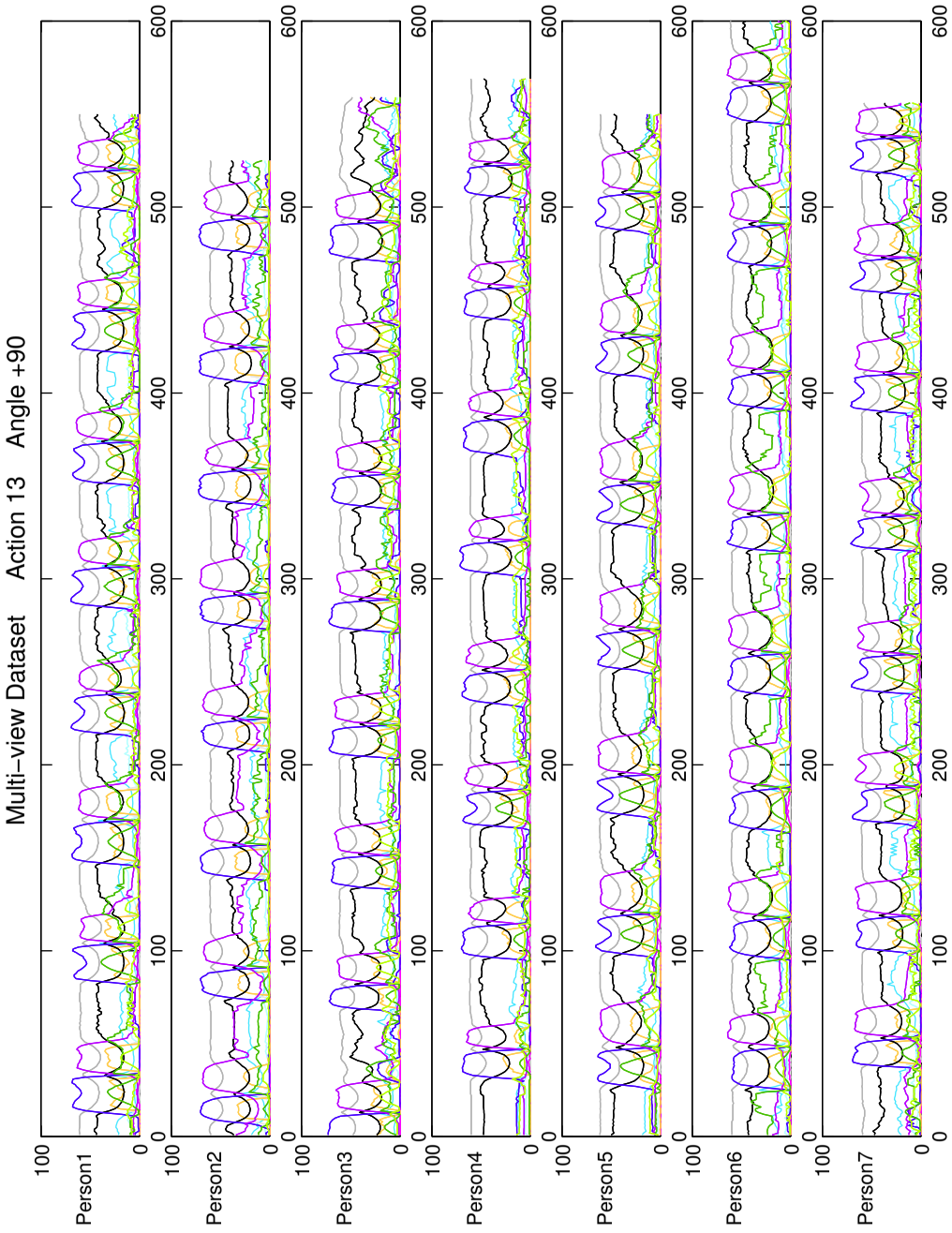

Figure 4.24: Feature traces from Action 1 viewed at an angle of 90 degrees.

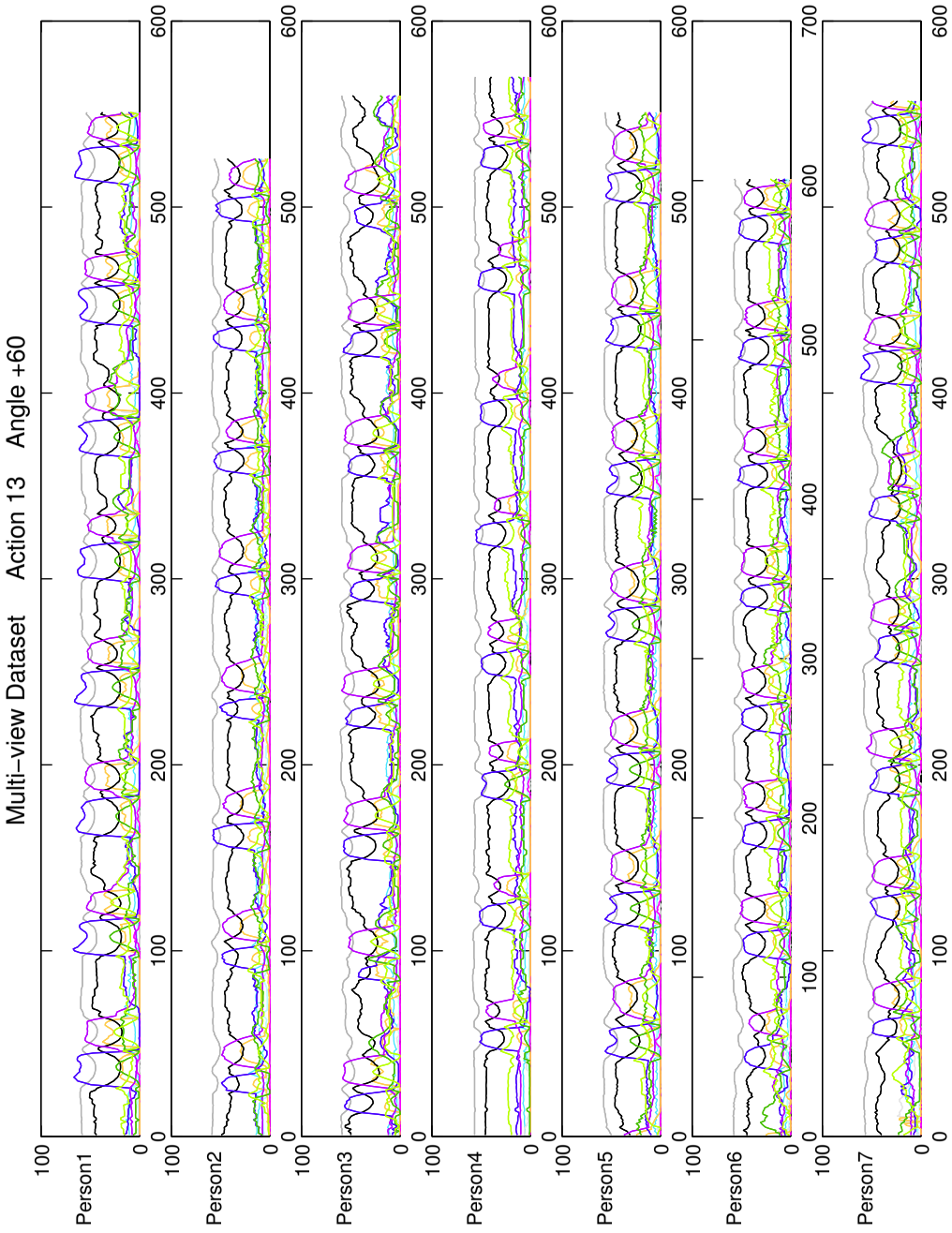

Figure 4.25: Feature traces from Action 1 viewed at an angle of 60 degrees.

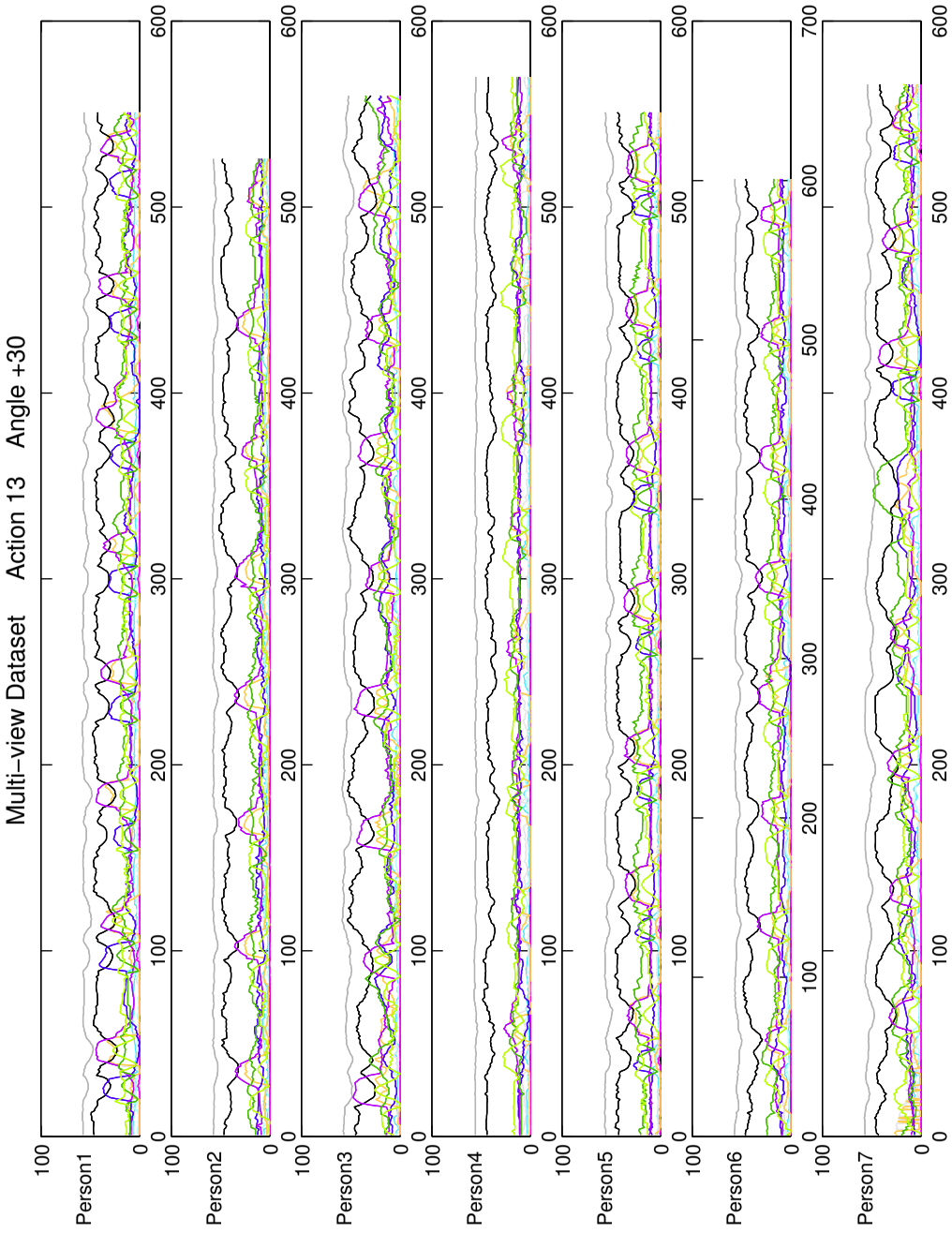

Figure 4.26: Feature traces from Action 1 viewed at an angle of 30 degrees.

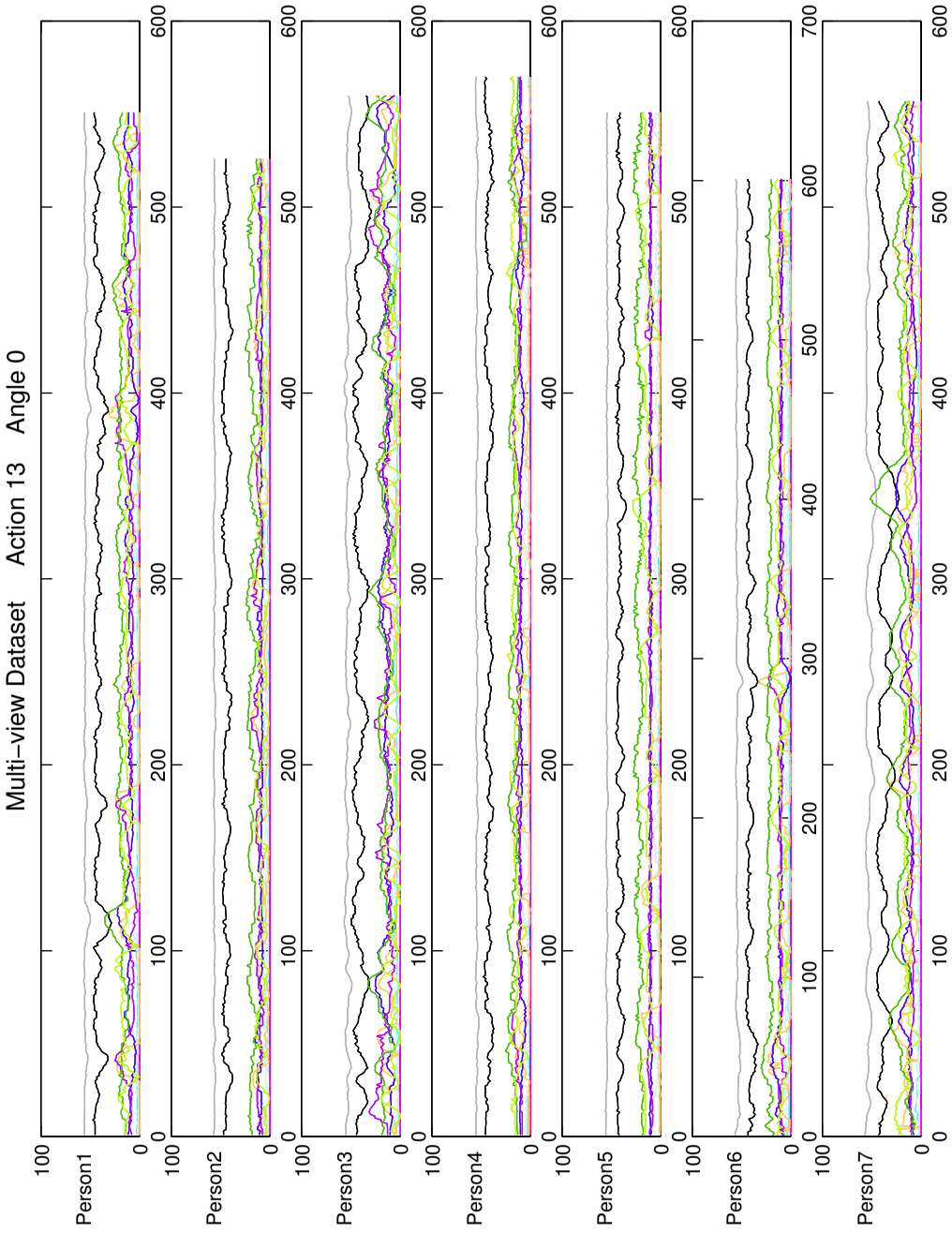

Figure 4.27: Feature traces from Action 2 viewed at an angle of 0 degrees.

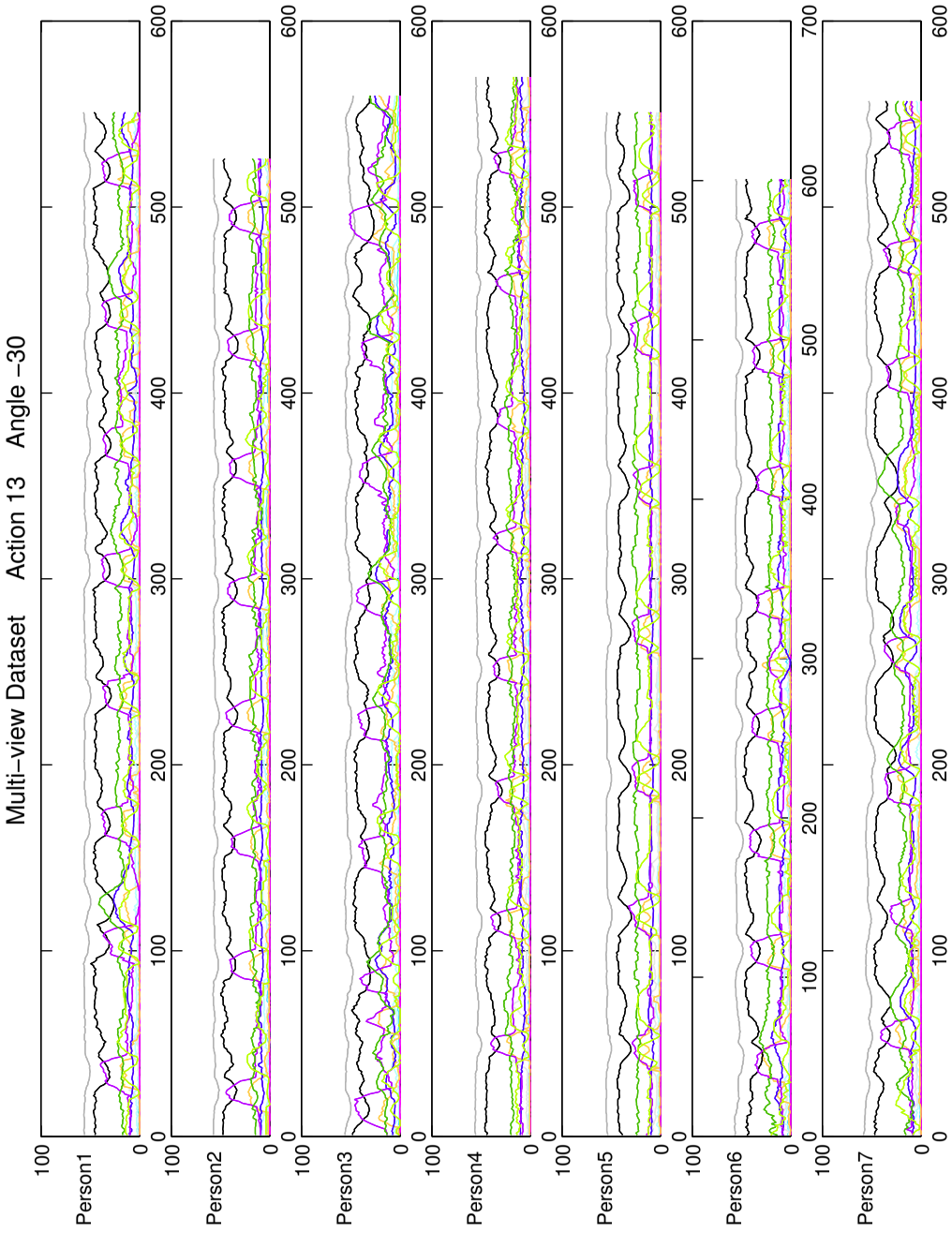

Figure 4.28: Feature traces from Action 2 viewed at an angle of -30 degrees..

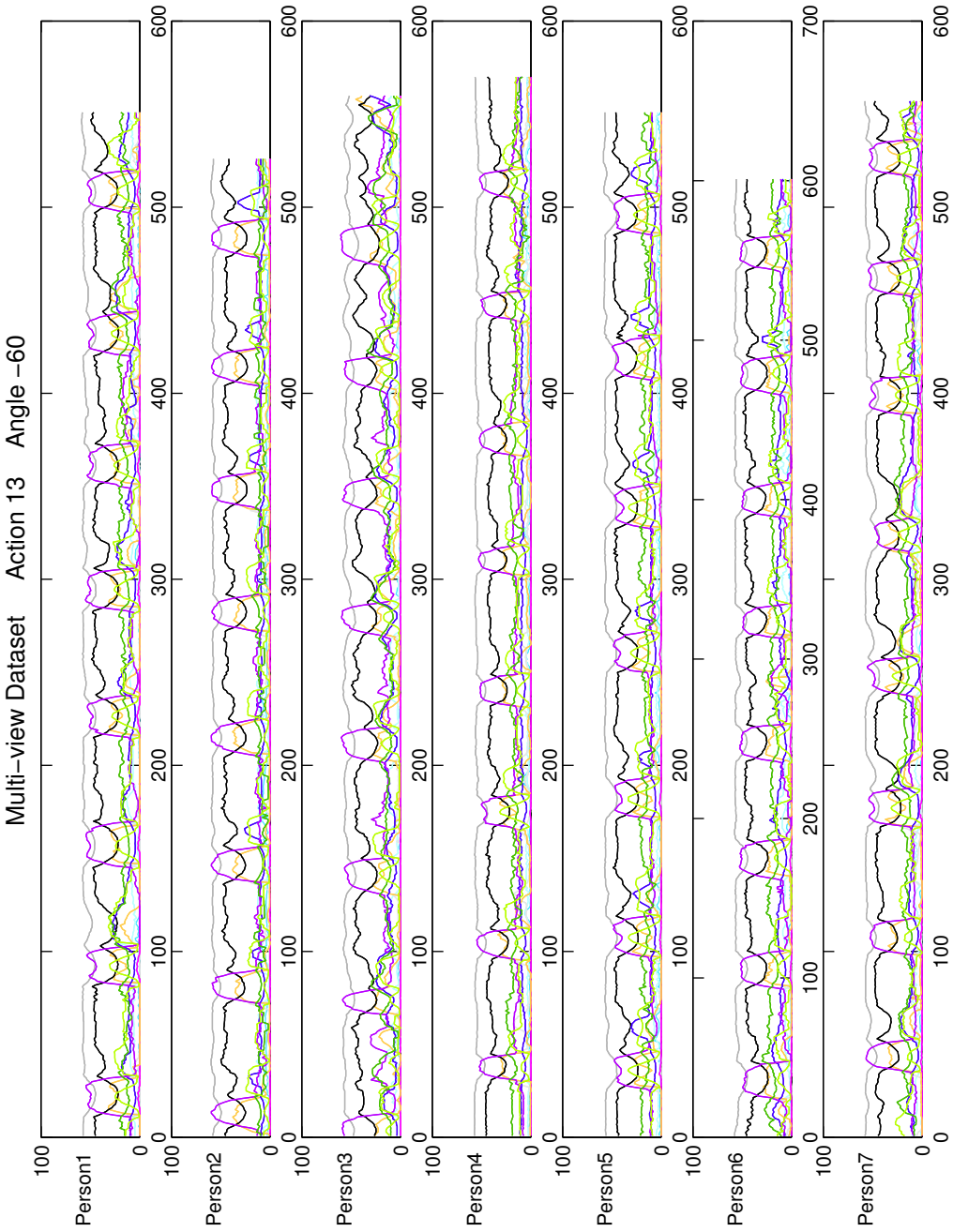

Figure 4.29: Feature traces from Action 1 viewed at an angle of -60 degrees.

Deviations within an action group cause difficulties when considering analysis methods such as Dynamic Time Warping that rely on characterising the actual trace shape. To achieve recognition such methods not only have to model the differences for each individual feature in the vector, but for the other 10 as well. This process would have to be repeated for each new action.

A study of the colour plots shows that the different actions can be characterised by the colour and bounding box features. Despite variations in these traces caused by time differences, personal appearance and mannerisms, a clear recognisable pattern can be associated with an action or different parts of an action. There are specific stages during the action where the traces correspond making the action recognisable. The following chapter describes how actions and poses can be recognised by exploiting these similarities between parts of the traces.

Chapter 5

Human Pose and Action Recognition

5.1 Processing the negative space features

Chapter 4 describes how features are extracted from the colour coded images. Each image in our datasets can now been represented by an 11 element feature vector that represents the negative space colour percentages and bounding box ratio of the image. This chapter discusses how a training set of the feature data is used to construct a pose classifier. The pose classifier is then used to reduce the information in an image to a single label indicating which pose class it belongs to. Action recognition then becomes the detection of a unique string of these pose labels in an input sequence.

The traces in Figure 5.1 below constitute the different negative space, colour coded feature traces over time of 13 different repetitions of a wave action some performed by different actors. This representation shows how difficult it is to generate a model that could be used to recognise the same actions. People perform an action in very distinctive ways and over different periods of time; a model of the action should be invariant to these differences. Although the traces are too complex and show much variation over time, they do hold the same relationships between one another. For example, it can be seen across each of the wave sequences that the relationships between each grey trace and the corresponding red trace evolve in a similar manner over the entire action.

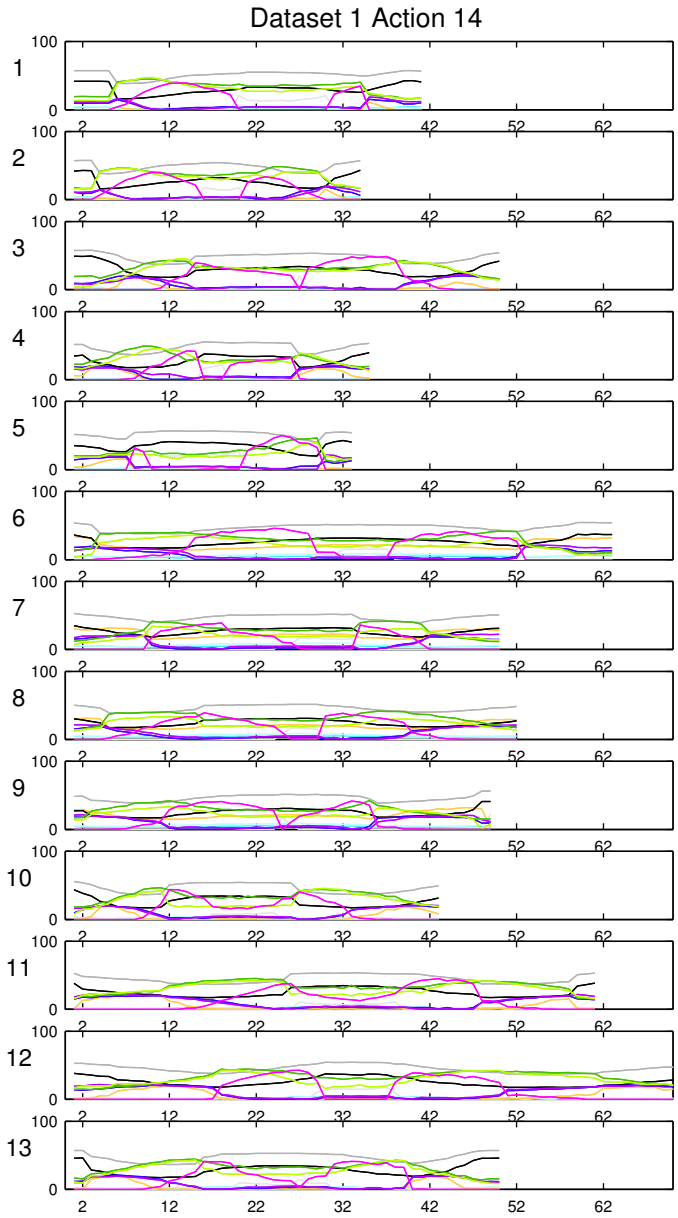

Figure 5.1: Traces of wave sequences from Dataset 1. These illustrate temporal and appearance related differences.

In other words, the traces may not have the same shape, but the relative proportions of the colours to one another do conform to a pattern. These relationships over small time periods imply that the bodies have gone through similar pose configurations. If we could group these, we could replace segments of the traces with the corresponding pose group. Thus, an action trace can then be represented by a sequence of characteristic poses as shown in Figure 5.2.

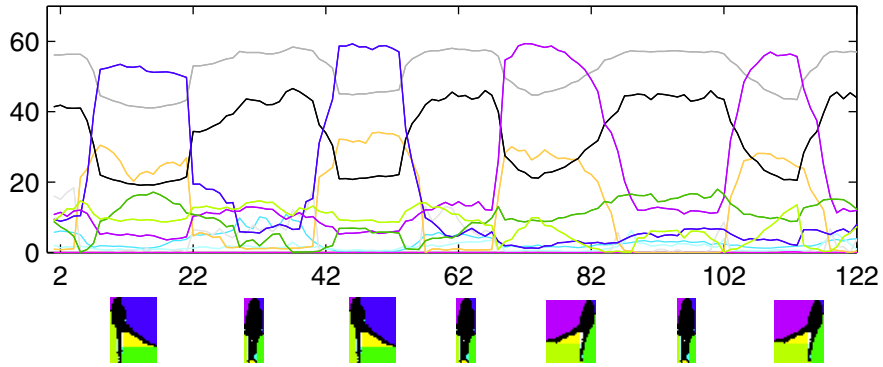

Figure 5.2: Selected traces from a kicking sequence. The images relate to the coloured traces above them. A sequence of these poses could compactly describe the action.

5.2 Pose classification

The pose approach leads to a simplification of the action recognition problem. Poses can be seen as letters in an alphabet: as long as you have the alphabet, new words can be made up through different arrangements of the letters. If you were to extend this analogy, just as any written language has a far greater number of words than alphabet characters, so to can this negative space 'language' represent a large set of actions using a concise set of poses.

The concept of an action built up out of blocks of simpler movements, is not new, and has been described in terms of movemes and phonemes [10]. Movemes are simple movements and phonemes describes actions made up of these movemes. These are still, however, movements over time. The work described here differs slightly in that each image in the sequence is itself assigned to a characteristic, or key pose, representing a range of similar poses [38]. A sequence of these characteristic static poses can now describe an action.

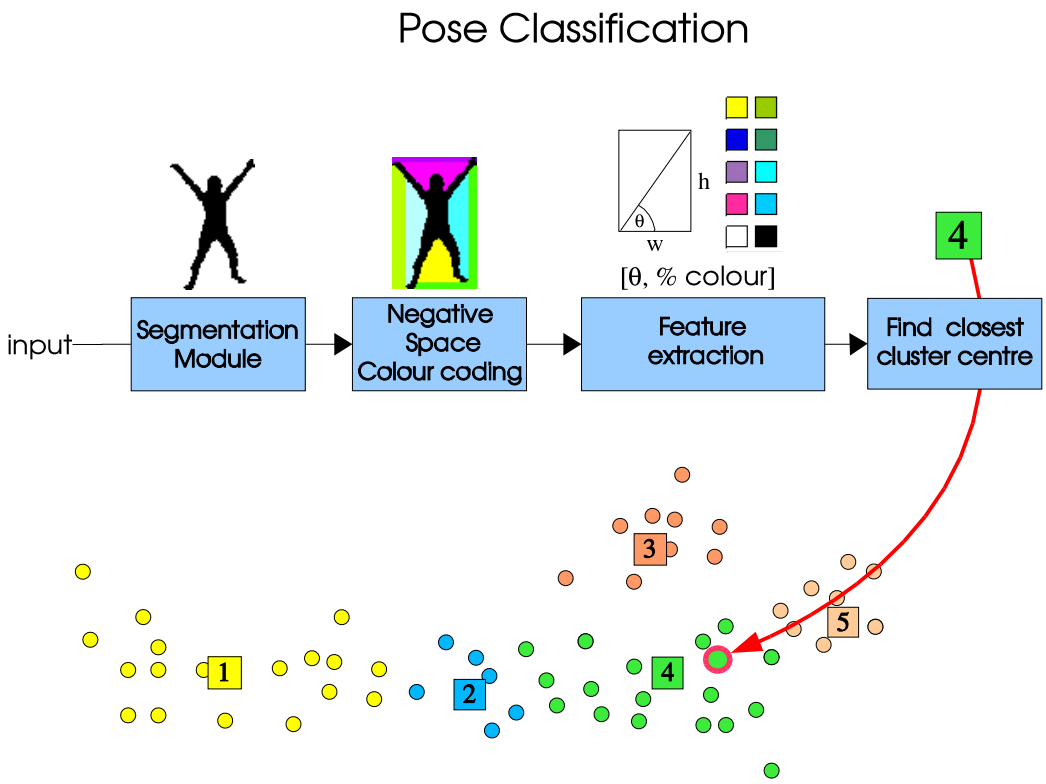

Figure 5.3: An image from an action sequence receive a pose label. This pose label corresponds to the cluster centre closest to the data point.

There are some practical considerations that also contribute to the preferred pose based approach. These include camera frame rates and the dropping of frames. To limit the size of the dataset, a trade off exists between the amount of actions that are filmed and the frame rate at which they are filmed. In the current system, the frame rate is not high enough to accurately capture a fast motion. This implies that the same motion could reveal dissimilar sequences depending on the time instances at which they were sampled. However, if only a single frame in a kicking sequence contained a key pose characteristic of the action there is still a good chance that it could be recognised. The pose-based approach is therefore preferred as it is more robust to a variety of different sampling conditions.

The first task is now to establish a set of characteristic key poses that can adequately categorise similar poses into effective groupings. This essentially implies a partitioning of the space containing the multi-dimensional feature data into regions demarcating similar poses. A pose label could then be assigned to an image based on which region its feature

data resides in. Manual thresholds for this demarcation could be set, but this would be uninformed, imposing a structure on to the data that may not be meaningful. A more suitable approach would be to use characteristics of the feature data to infer effective pose groups from the data structure itself.

The number of regions in the data space indicates how general the pose grouping is; the size of the regions must be such that we group like poses and but not so general as to group entire actions or so specific as to differentiate based on personal appearance. The number of regions is also dependent on the actions performed; the more uniquely distinct actions, the more poses and the more varied the traces. More regions would have to be demarcated to accurately segment the pose groups.

5.2.1 Creating pose regions using *k*-Means clustering

The reliance on inherent data structures strongly suggests the use of clustering methods to partition the feature data. There are a number of available methods to cluster data [20, 36, 74] ranging from supervised to unsupervised clustering algorithms. In supervised methods the category labels are available whereas in unsupervised methods they are not.

One of the most popular and simple unsupervised clustering methods is the *k*_Means clustering algorithm [20, 37]. This approach divides the input data into a specified number of clusters. The goal of this procedure is to find the c number of cluster centres μ_1, μ_2, ..., μ_c by computing the squared Euclidean distance of each data point to each possible centre $\| x_k - \hat{\mu}_i \|^2$ and then finding the mean $\hat{\mu}_m$ nearest to x_k . If the number of data samples is *n* and the desired number of clusters *is c* then the algorithm is as follows [20]:

begin initialise n, c, μ_1, μ_2, ..., μ_c

do *classify* n *samples according to nearest* μ_i

recompute μ_i

until no change in μ_i

return μ_1,μ_2, ..., μ

end

The procedure can be explained as follows:

Take K data points at random from the dataset. Each of these seeds are used as cluster centres and we assign the rest of the data points to the closest centre. Every data point in the set now belongs to a cluster. The centroids of these clusters are then calculated as the average position of all the data points in this cluster. The data points are again assigned to the closest centroid. This process of assigning the points to clusters and recalculating the centroids are continued until the cluster boundaries stop changing.

5.2.2 Difficulties with clustering the features

If a training set containing a good representation of the feature data is clustered into its pose representation it would be possible to automatically identify the pose of a new image by associating its feature vector with the closest cluster centre. There are, however, two factors complicating the design of such a system; the poses do not form well separated clusters and there is no labelled training set.

A labelled training set would simplify the task of pose partitioning; the number of different poses would be known beforehand as well as how they are clustered across the data space. However, it is an impossible task to manually label these features with no idea as to what the pose partitioning should be. Well-separated clusters allow for a simple, geometric based classification. Unfortunately, the pose feature data does not manifest well separated clusters. Upon reflection, this is necessarily the case because the data is derived from sampling a continuously changing entity. Simply put, performing an action is a fluid motion; the poses gradually change from one to another where the only discontinuities are colours disappearing or appearing. This assumption is substantiated by the following self organising map (SOM) [2], which allows for the visualisation of the relative distances in the high dimensional space. Areas with colours higher up on the right hand scale indicate regions of large relative distances between data points. The figure thus confirms that there are discontinuities amongst the data but an overall lack of well separated clusters.

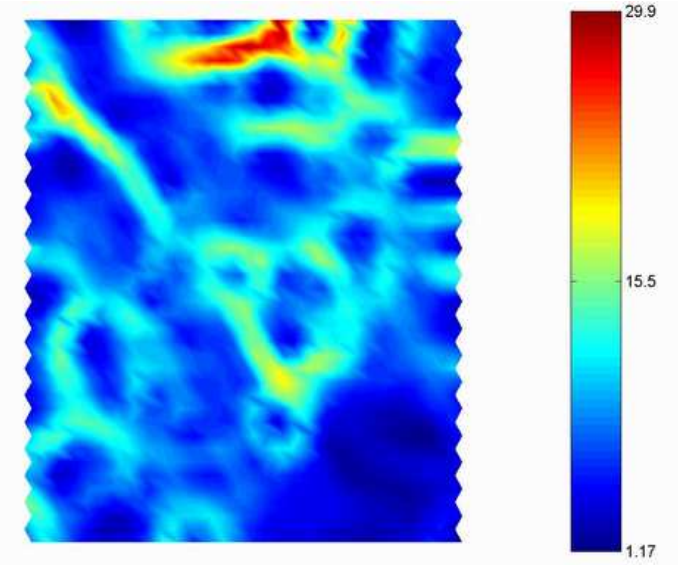

Figure 5.4: SOM map of feature data showing the separations in the data, the scale on the right show how the colours indicate relative distances in the data. The higher up the coloured area on the scale the further apart the data in that area.

Further support can be found from independent cluster validity methods. These methods are used in colour segmentation problems, for example, to determine the ideal number of clusters. Validity methods include the Davis-Bouldin and Dunn indices [19]. A validity measure most applicable to the pose data, assuming that clusters were present, can be found in the work by Turi and Ray [77]. To find the best clusters they propose that one minimise:

$$validity = y \times \frac{intra}{inter} \tag{5.1}$$

Where intra refers to the average distance to the cluster centre for points within the cluster. Inter is the minimum distance between all the different clusters. Y is a function of the number of clusters [77], and is used to penalise a small amount of cluster centres, to avoid overgeneralisation. This validity measure is supposedly better than the Davis-Bouldin and Dunn indices but it does not converge for the pose data, supporting the discontinuity, but no-cluster theory.

5.2.3 Clustering using feedback

It has been established that there are no natural cluster groupings and there are no manual ways to label the pose classes. Consequently, unsupervised clustering methods fail to converge on a unique solution set as the feature data does not form distinct clusters. Furthermore, supervised clustering methods require the knowledge of the number of clusters expected as well as a labelled training set, both of which are unavailable.

Suppose, however, that hard borders are imposed at intervals within the data space, this would in effect quantise the negative space features, forcing pose partitions on the data. With an appropriate partitioning, the actions will be divided up into pose representations; a tour through a specific sequence of pose regions would then describe a particular action. The decision of where to place these borders is critical: too wide a sectioning could capture all poses within an entire action whereas too small a sectioning could effectively put each individual data point into its own pose class. Stating it conversely, the partitioning has to be general enough to allow for variabilities that arise due to appearance and mannerism, yet small enough to adequately distinguish different actions. Therefore there must exist an optimum partitioning of the feature data which can be evaluated by how the action traces are segmented.

Although there are no labelled poses, it is possible to label different actions. Conceptually each different action group should be made up of a tour through its unique sequence of characteristic pose partitions. Thus, although we do not know beforehand what the partitioning of the feature space should be, similar actions should ideally be comprised of similar sequences of poses. This similarity can be used as a fitness measure when comparing the different ways of partitioning the data.

In order to test different numbers of partitionings a data set must be divided into a training set, a feedback set and a testing set. The training set is that set of data to which the k-means algorithm is applied. This set must be constituted of a wide variety of poses to obtain the most comprehensive classifier possible.

The feedback (or validation) set is that set of data used to measure the fitness of each outcome of the k-means algorithm. This set is comprised of groups of known action sequences each performed by different actors, thus encapsulating the variability that can

arise when an action is performed by different people.

The testing set is an independent data set used to test the performance of the system.

Figures 5.5, 5.6 and 5.7 illustrate the results of different partitionings on the action sets. The images merely illustrate the ideas, the actual feature data exists in multi-dimensional space and cannot be visualised.

Suppose that the negative space features from a training dataset have been mapped into a number of different pose groups, then Figure 5.3 on page 54 shows how an image from one of the action sequences is thus assigned. The image is now represented only by this pose group label. A sequence of images of a person performing an action will then be replaced by a sequence of numbers corresponding to the pose groups the images have been assigned to. As the consecutive images contain very similar poses these are very likely to have been assigned to the same pose group. Thus an action sequence can contain series of duplicate numbers for those consecutive frames.

If a number of the sequences of people performing the same actions are labelled in this manner, it becomes possible to compare different ways of partitioning the data.

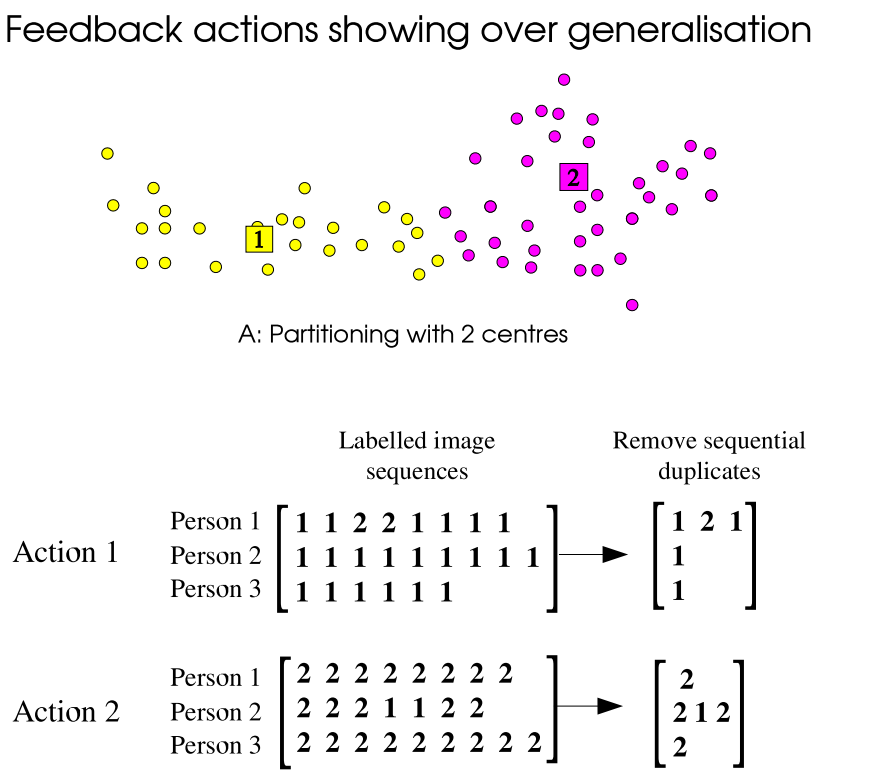

Figure 5.5: Hypothetical illustration of the effect of using too few pose clusters

Figure 5.5 shows an example where a comparison of the labels of the feedback action sets show that the data has been partitioned into too few sections. The feature data space has been divided into two sections, thus a pose can belong to only one of two groups. If every image in an action sequence happen to belong to pose label 1 then an entire action will be classed as one pose. Repeated sequential numbers are removed as they represent a tour through the same pose partition. After removal of these duplicates the sequence of pose labels that make up an action is easier to identify.

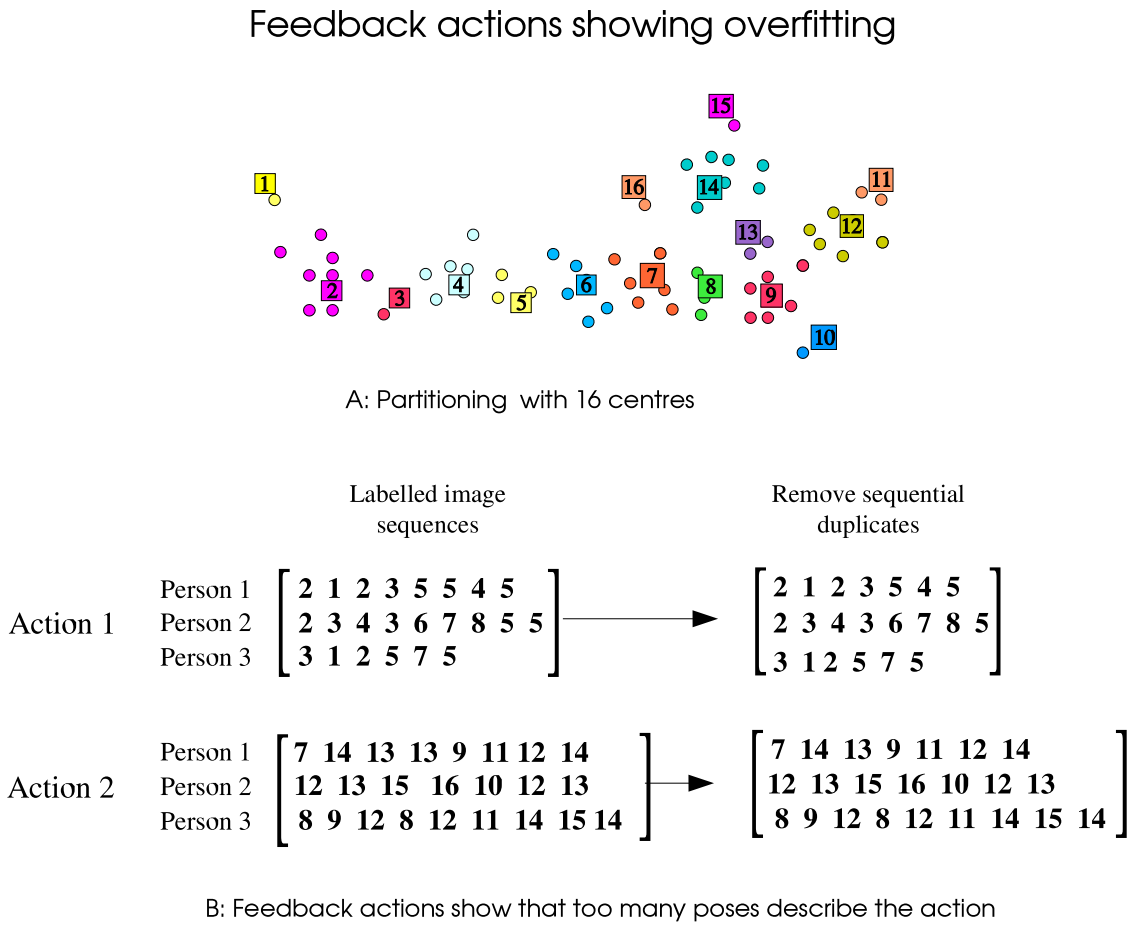

Figure 5.6: Hypothetical illustration of the effect of using too many pose clusters

Figure 5.6 shows what would happen if there were too many partitions; the system starts to distinguish pose regions based on personal appearance and small deviations in the manner which different people perform the same action. In the most extreme case all the images will be assigned to a different pose label and none of the action sequences would contain the same numbers let alone the same sequence of numbers.

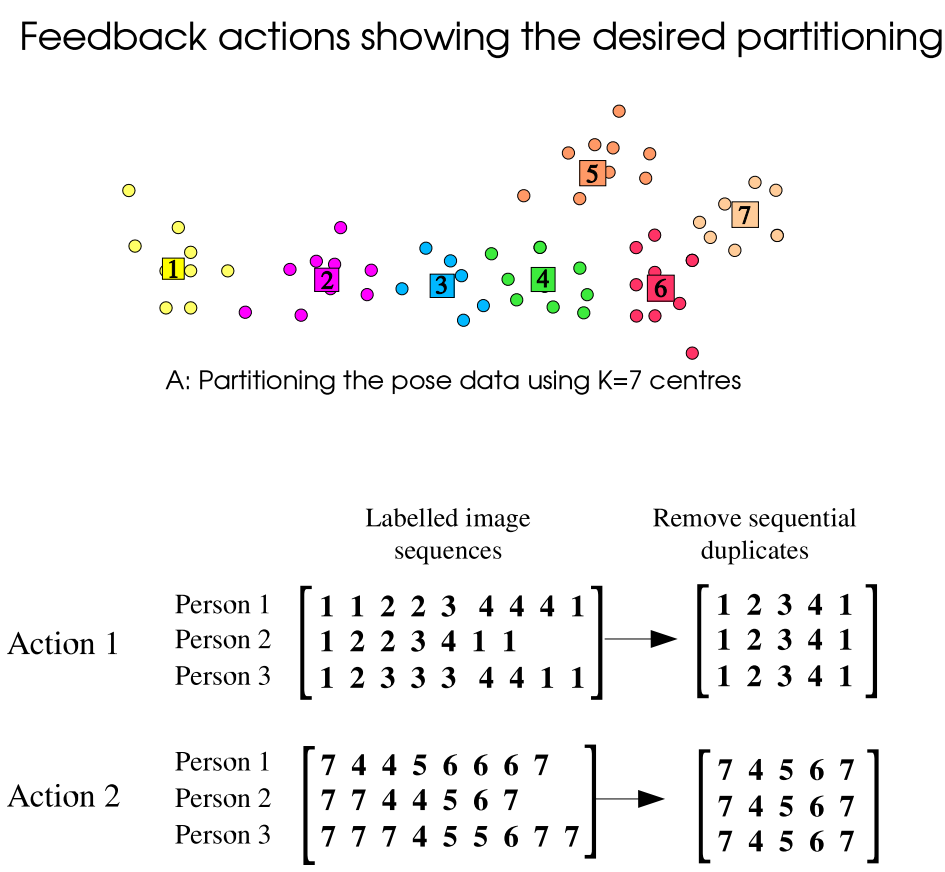

Figure 5.7: Hypothetical illustration of the desired pose partitioning

The desired partitioning is illustrated in Figure 5.7. Here the data is partitioned into regions that capture the same underlying characteristics. Thus, when the feedback actions are segmented and labelled they contain the same sequence of poses. Because not all people do the same action over the same length of time, some people would remain longer in a particular region of feature space. This accounts for repeated pose labels in the different labelled image sequences. Once these duplicates are removed, the actions are all a tour through the same sequence of pose regions. The same action performed in another sequence can now be recognised by simply detecting the numbered string corresponding to that action.

In practise, the partitioning as illustrated in Figure 5.7 would be very rare. Instead it is more likely that a given partitioning strategy groups some of the features optimally, while others are divided into neighbouring regions that will then contain poses that are all very

similar. These similar neighbouring regions generate different labels for feature data that should have ideally been classed into the same pose group. Even if the optimal partitioning could be found there might still be neighbouring regions containing practically similar poses making the recognition of the action more difficult. If these regions were merged, however, would provide feedback action sequences with segments that are labelled even more alike for the same action.

By extracting features from the negative space images , we have lost a lot of information in the more convenient representation. However, this information is still available because it is possible to associate an image from the training set with the corresponding negative space feature data point. Every cluster centre can also be associated with an image simply by assigning it the image associated with the data point nearest to it. Figure 5.8 shows conceptually how each cluster centre can be assigned a colour coded image. These images provide additional information on the individual partitions. It now becomes a matter of image correlation to decide whether partitions should be merged. Proximate clusters can be merged if their correlation values are higher than a set threshold value.

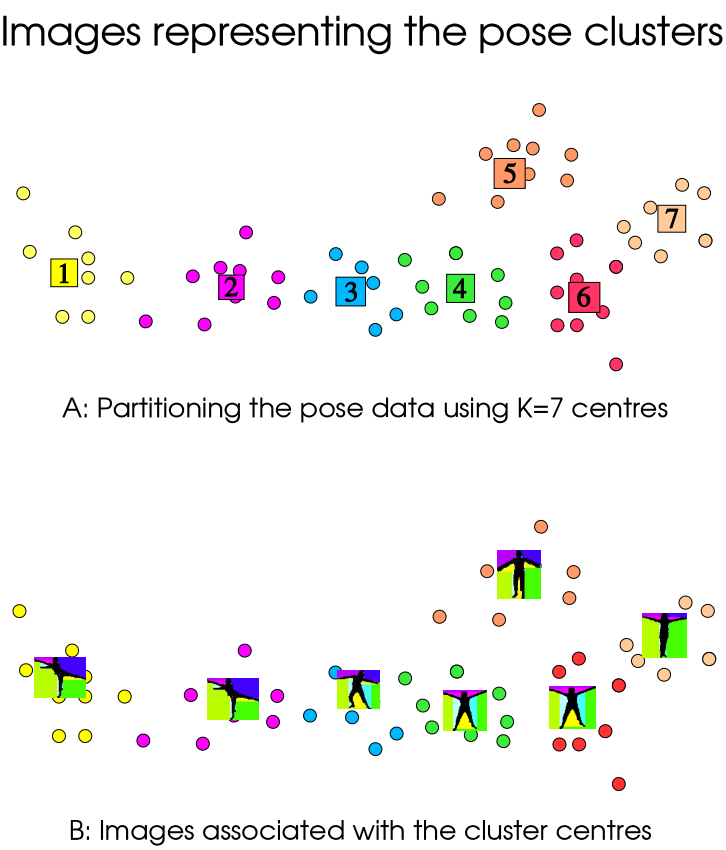

Figure 5.8: Image A shows the partitioned feature data. Image B shows images taken from data points closest to the cluster centres.

The next section describes how to overcome the difficulties discussed in this section and to automate the task of constructing a pose classifier using the feedback set to determine both the number of clusters K that the data should be partitioned into, and T the correlation threshold for which proximate clusters should be linked.

5.2.4 Automatic partitioning of the pose data using feedback

This section describes how the *k*-Means clustering algorithm and image correlation are iteratively used to determine the best number of partitions that the feature data should be sectioned into. The algorithm is related here in pseudo code, each of the steps is then discussed in more detail.

1. **for** K=2:number of images in training set
   - Cluster data into K centres using *k*-Means
   - Determine fitness of clustering using feedback action set

   end
2. Compare fitnesses to find optimal K.
3. Use the best partitioning as found in 2 and find the image associated with each cluster centre.
4. **for** T=0.6 : 0.01 : 0.8
   - Evaluate the correlation of all neighbouring cluster centre images
   - Merge clusters if their correlation is bigger than T, and relabel the clusters accordingly
   - Determine fitness of new merged clusters on feedback action set

   end
5. Compare fitnesses to find optimal T.

The algorithm operates in mainly two parts. Firstly the optimal number of cluster centres, K, is determined by evaluating the fitness of a number of clusterings against the action feedback set. Secondly, using the value found for optimal K, a number of correlation thresholds are used to merge clusters if the images representing neighbouring centres result in a correlation value higher than the threshold T. The optimal correlation value is again based on results from the feedback set.

For each value of K the training set is clustered using the *k*-Means algorithm. The resulting partitioning is then used to label the poses in the feedback set. This pose label is the cluster number of the centre closest to each data point in the feedback set; a simple Euclidean distance measure is used to determine this closest centre. This is done for each value of K. The optimal number of clusters, optimal K, can be found by evaluating the labelled feedback sequences. This evaluation is done before correlation merging and the feedback set will not exhibit a large amount of similar sequences. At this stage, this does not matter as the interest lies only in the optimal value for K . Figure 5.9 illustrates the

process.

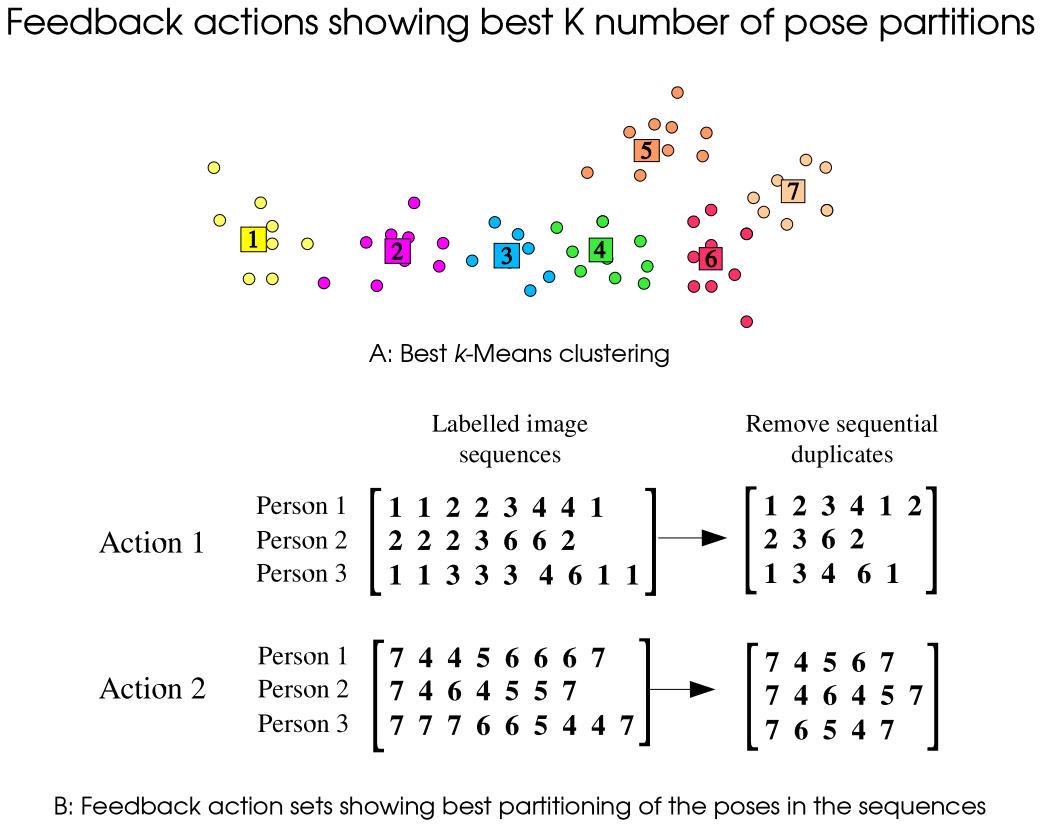

Figure 5.9: The clustering process illustrated in 2 dimensions for clarity. Image A a shows the feature data clustered for the optimal K as determined by the feedback action sets shown in B.

Two different criteria are used in order to compare the pose partitioning on the feedback actions, a longest common sequence measure and a penalty measure.

The longest common sequence measure compares the pose sequences within each action group to one another seeking the longest common sequence of pose labels. This sequence need not be continuous over the action, it just has to follow the same order. For example, consider the segmented actions with pose labels [1 2 3] and [1 2 4 3 1], the longest common sequence has a length of 3 [1 2 3]. Notice that the second '1' is not considered as [2 3 1], another possible sequence containing the same poses does not conform, to the sequence order.

Consider a second comparison between sequence [1 2 4 3 1] and sequence [6 1 4 7 2 8 4 9 1 3]. Again the longest common sequence is [1 2 3], however clearly the second sequence is

unacceptably too far from the longest common sequence. Therefore a measure of how similar the feedback sequences are to the longest common sequence among them needs to be calculated.

To do this, the common sequence length (in this case a length of 3) is then divided by the average length of the pose sequences. A set of completely identical sequences would carry a measure of 1 whereas a set of completely different sequences would carry a measure of 0. Thus, this measure indicates how similar actions are to one another within the group. This fitness is calculated for each group of like actions. The overall longest common sequence fitness over the whole feedback set is simply then the average of these.

The final criteria is a penalty measure; no series is allowed to have a length equal to or smaller than 2 pose labels (remembering that sequential duplicates have been removed) as this results in over-generalisation.

At this point the optimal number of clusters, optimal K, to partition the data into has been determined. However, this partitioning could have a number neighbouring clusters that contain data extracted from images that are very similar. Image correlation is now used to establish which of these clusters should be merged. Clusters are merged by assigning the same label to neighbouring clusters that derive from similar images. The image correlation stage requires a colour coded image to be associated with each individual cluster centre. The feature data is searched to find the point closest to the cluster centre and the colour-coded image corresponding to this data point is then used to represent the cluster in the correlation calculations. Figure 5.10 illustrates how clusters are merged if they have a correlation value higher than T.

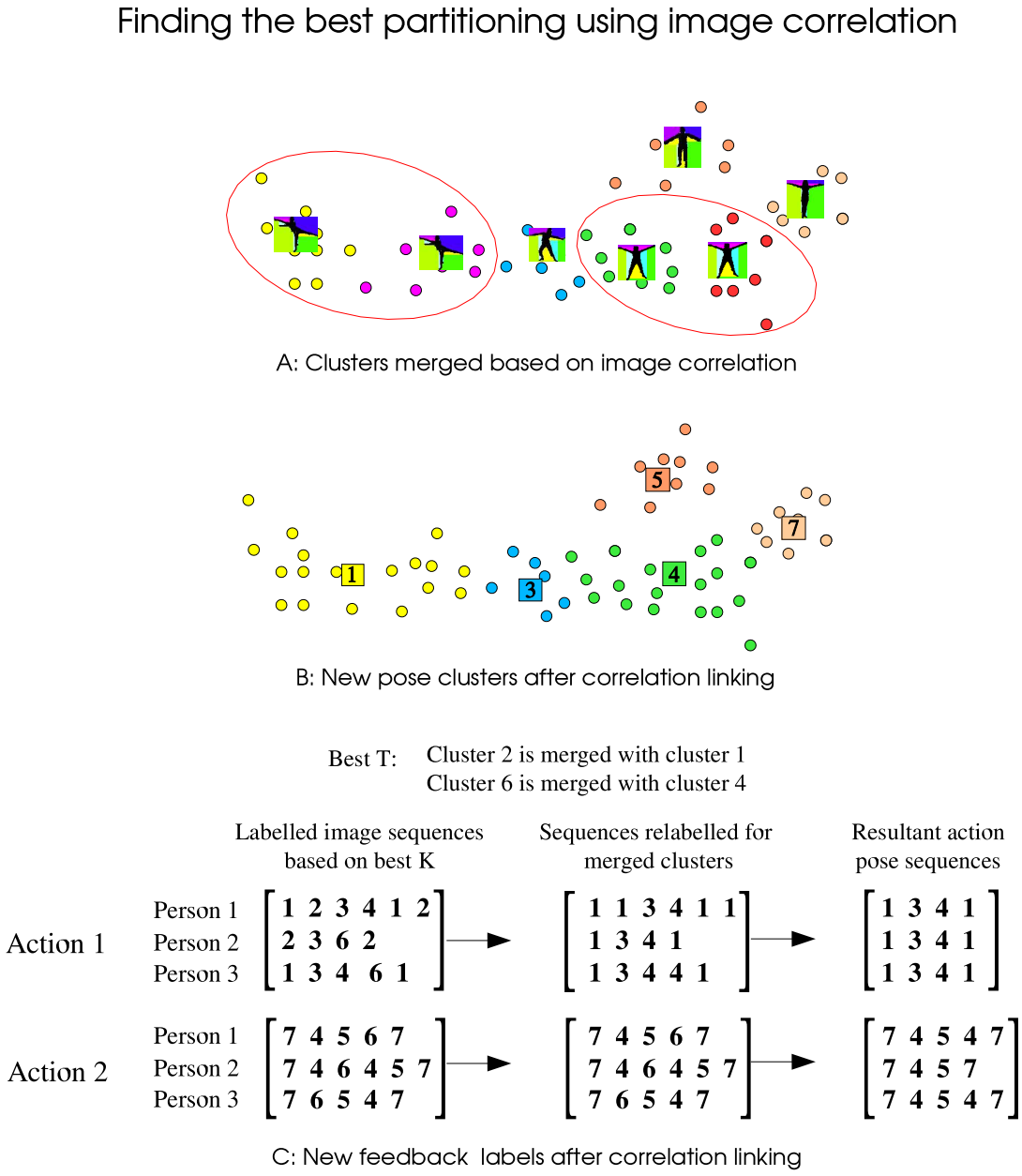

Figure 5.10: Neighbouring partitions that are too similar are merged if the correlation of the image representing the cluster centres are high enough.

Close cluster neighbours alone are considered as these images will suffer only in a minor sense from the effects of distortion that occur when the images are normalised for 2 dimensional correlation for they will have almost the same dimensions.

These image matrices A and B are normalised for comparison, the normalised cross-correlation coefficient is calculated using the following method, m and n are the dimensions of the normalised images

$$r = \frac{\sum_m \sum_n \left(A_{mn} - \bar{A}\right)\left(B_{mn} - \bar{B}\right)}{\sqrt{\left(\sum_m \sum_n \left(A_{mn} - \bar{A}\right)^2\right)\left(\sum_m \sum_n \left(B_{mn} - \bar{B}\right)^2\right)}} \tag{5.2}$$

where

$$\bar{A} = \frac{1}{m * n} \sum_m \sum_n A_{mn} \quad and \quad \bar{B} = \frac{1}{m * n} \sum_m \sum_n B_{mn}$$

Correlation thresholds between T = 0.6 to 0.8 in intervals of .01 are used, previous analysis has shown that numbers smaller than 0.6 can produce pose clusters that are too big, in other words an entire sit up and lie down sequence would be seen as one pose which is not desirable. Thresholds bigger than 0.8 result in no alteration to the cluster configuration. This correlation was done only between binary silhouette images, although the correlation between the different coloured areas could also be considered.

The feedback actions are then again labelled for every value of the correlation threshold, T. This will result in a more compact labelling than was achieved using only optimal K as some of the clusters will now be merged with proximate neighbours. By using the same fitness criteria, the longest common sequence measure and the penalty measure, the optimal value for T, the correlation threshold, can be found to merge the neighbouring clusters.

By using K-Means to cluster the data, linking clusters based on correlation and evaluating the resulting segmentation on the feedback action sets, pose partitions are obtained that would be very difficult to achieve with unsupervised or supervised clustering methods that do not include knowledge of where the data originated from. The following sections relate how the algorithm was implemented on the datasets.

5.2.5 Partitioning the datasets

The algorithm in section 5.2.4 has been used to construct a classifier for the data from both Dataset 1 and 2. Two different classifiers were constructed. Dataset 1 consists of sequences captured in the university Chroma-key studio and is used to investigate the

tolerance of the system to variation in appearance. The size of the room allowed space for a single camera angle, making it not possible to view the same action from different angles. Dataset 2 was filmed in a large private film studio and contains sequences captured from several angles simultaneously to investigate the tolerance of the system to camera angle variation.

Dataset 1

Dataset 1 is of interest as the training set is made up of action sequences that are very different in appearance to those of both the feedback action sets and testing sets. The training set contains only women performing a wide range of free form actions, whereas the testing and feedback sets contain 2 men and 2 women wearing different clothing. By using the method outlined in the previous section, a training set comprising 1135 images was partitioned into optimal K=180 clusters. After correlation merging using an optimal threshold value of 0.7 these were further reduced to 83 pose regions.

Dataset 2

The pose cluster representation from Dataset 1 was not trained on multi-view data, and could not be expected to perform very well on Dataset 2 as the requisite poses are not present. A different training set is used for Dataset 2. This set is comprised of the entire set used in Dataset 1 as well as one repetition of an action, captured by each of the different available view angles in Dataset 2, for any two of the actors. This resulted in a training set of 4 879 images.

The resulting number of optimal partitions, optimal K is found to be 395 and the optimal correlation threshold value of 0.7 reduced the number of cluster centres to 195. The number of pose partitions are higher than that of dataset 1 as there are more varied poses present.

Figure 5.11 shows the images that represent cluster centres of the 395 clusters that result from the k-means procedure. Figure 5.12 shows the reduced 196 images after correlation linking of neighbouring clusters.

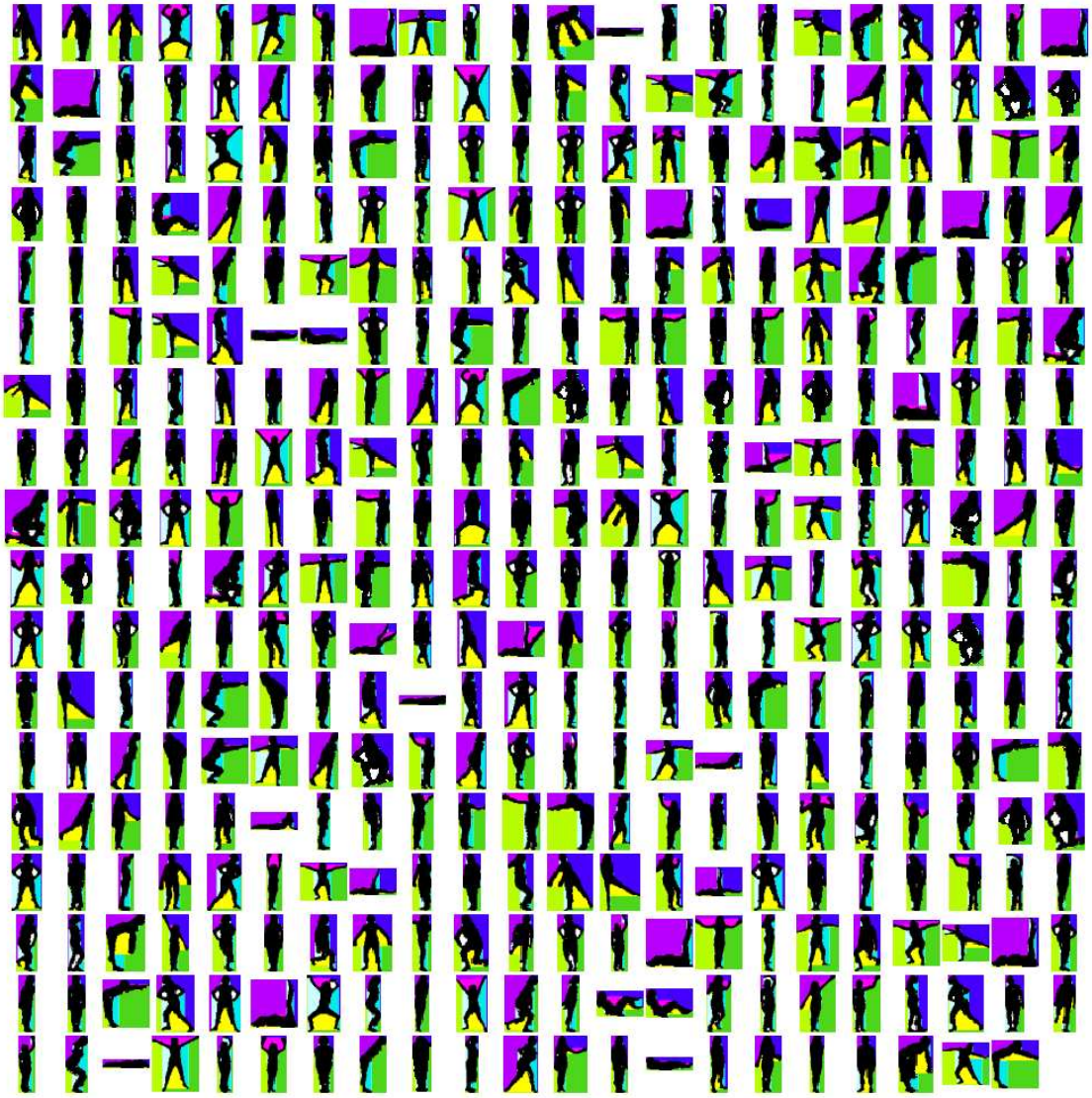

Figure 5.11: Images representing the cluster centres for Dataset 2

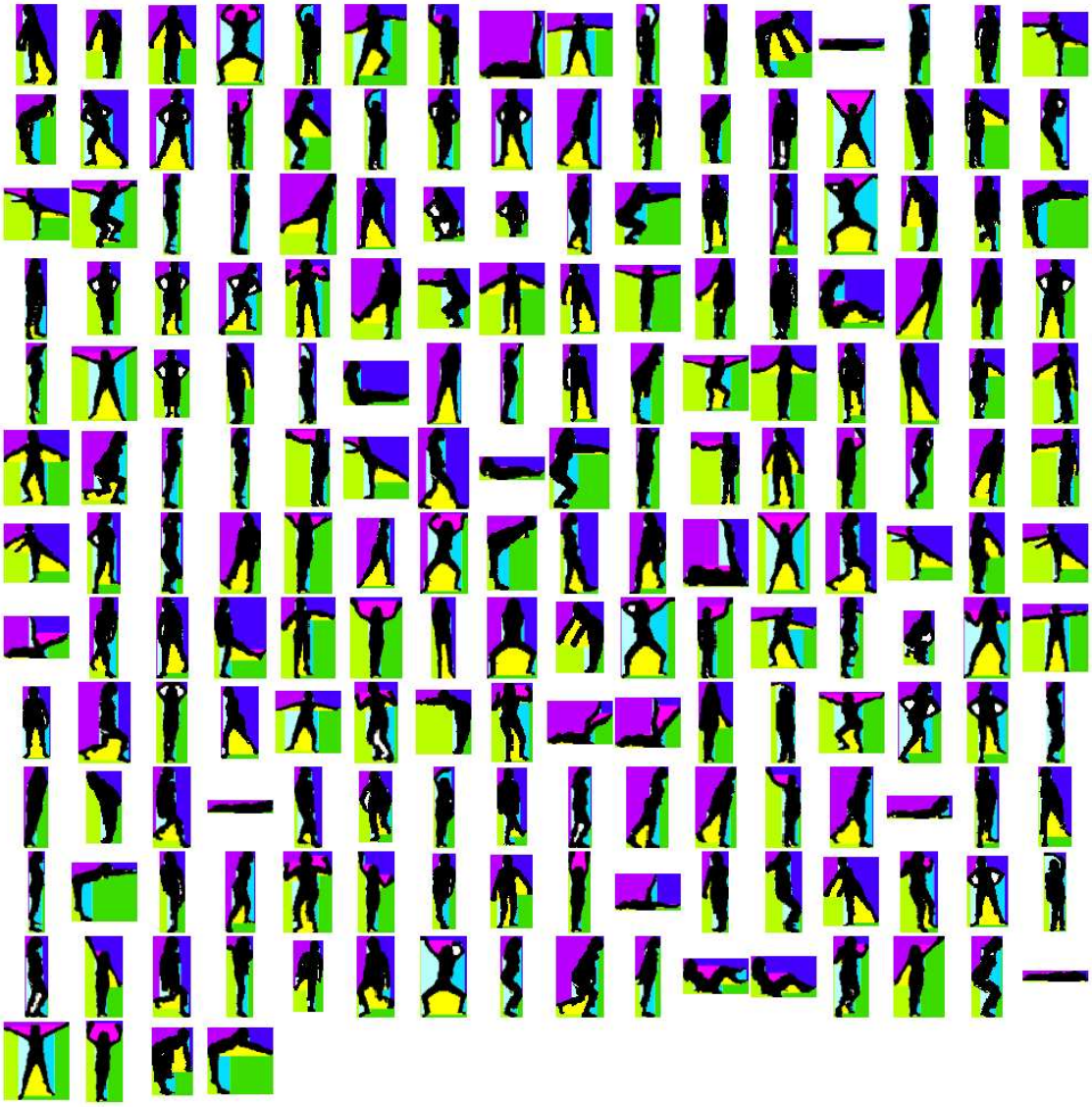

Figure 5.12: Image representing the cluster centres for Dataset 2 after correlation linking.

5.2.6 Benefits of the clustering approach

An investigation of the feature data plots in Chapters 4 and Appendix A and B have shown that actions that would generally be considered the same can appear differently in the feature space. These differences are due to personal appearance, manner of performing an action, segmentation noise, temporal differences and sampling differences due to the frame rate. Dissimilarity introduced by these factors create difficulties for methods that attempt to recognise the action signal in its entirety. The pose partitioning approach

described in this chapter overcomes many of these problems.

Figures 5.13 to 5.16 demonstrate the effectiveness of using the pose partitioning approach. The original trace is plotted in Figures 5.13 and 5.15 followed directly by their pose partition representations. The pose representation is obtained by plotting the cluster centre that a data point is assigned to instead of the data point itself. These traces along with those in Appendix C show visually the effectiveness of the pose partitioning because they are significantly less quantised than the raw data.

This section has described how to design a pose classifier for the negative space feature data. The following section explains how the classifier is used to recognise actions.

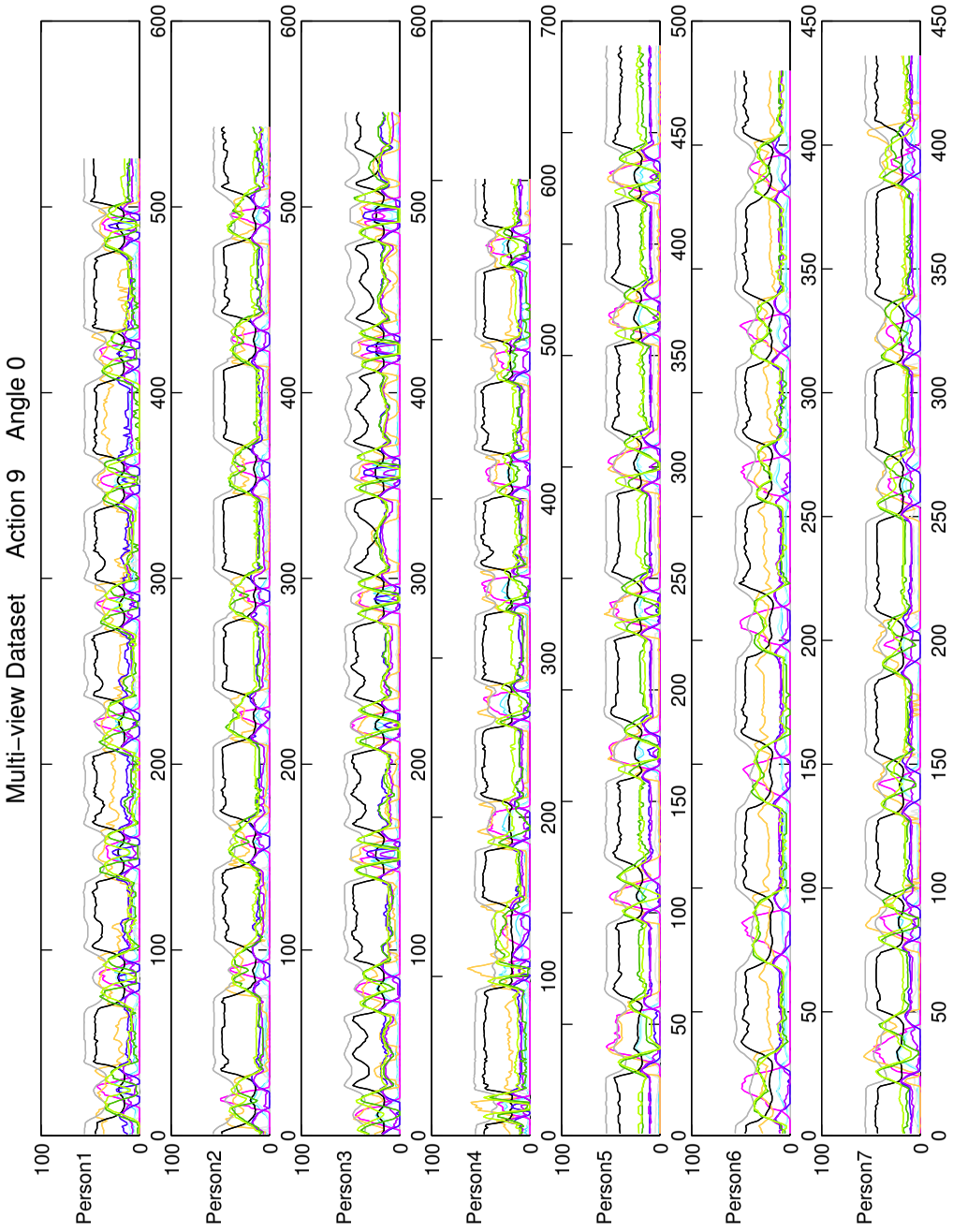

Figure 5.13: Actions from Dataset 2 before clustering.

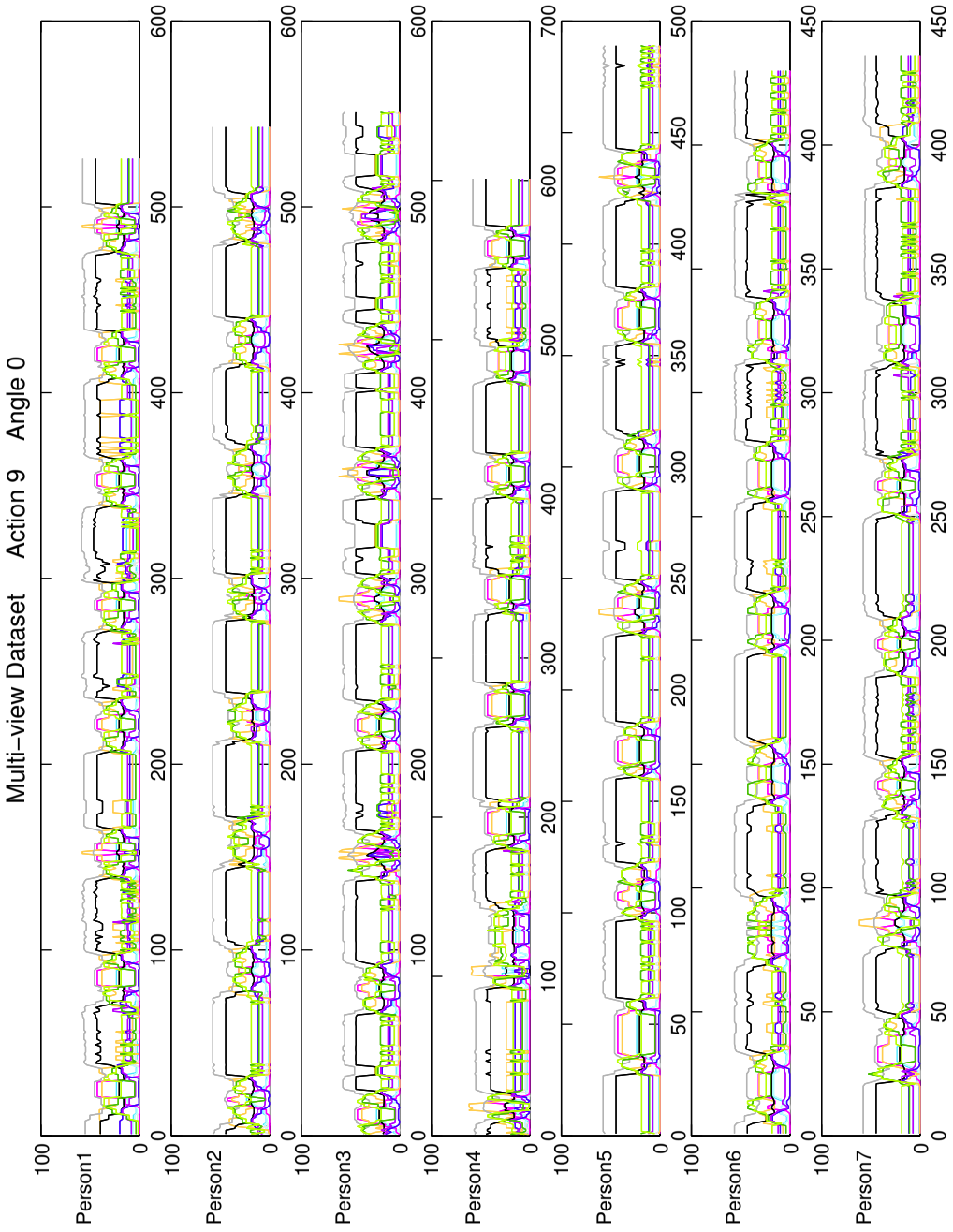

Figure 5.14: Actions from Dataset 2 after clustering

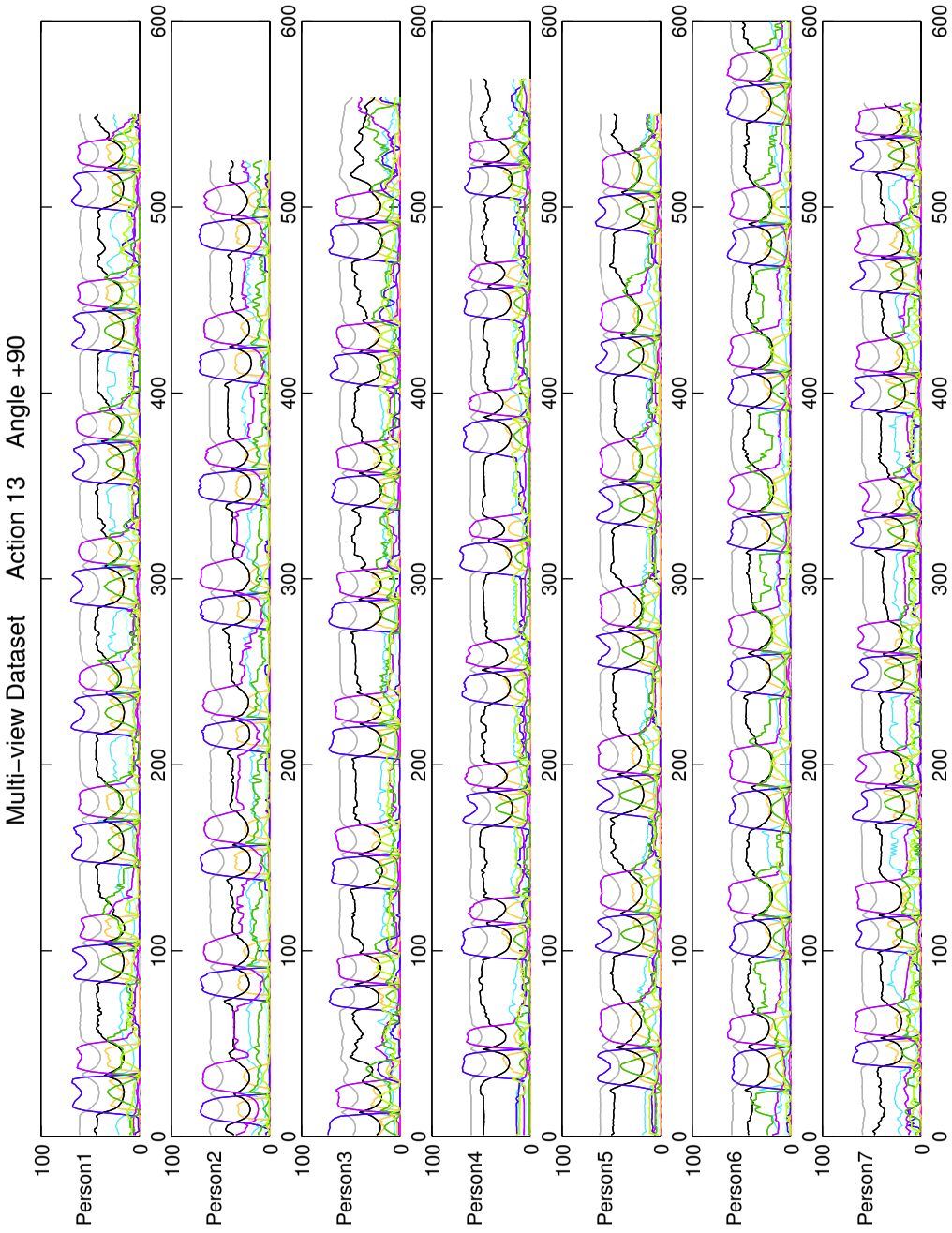

Figure 5.15: Actions from Dataset 2 before clustering.

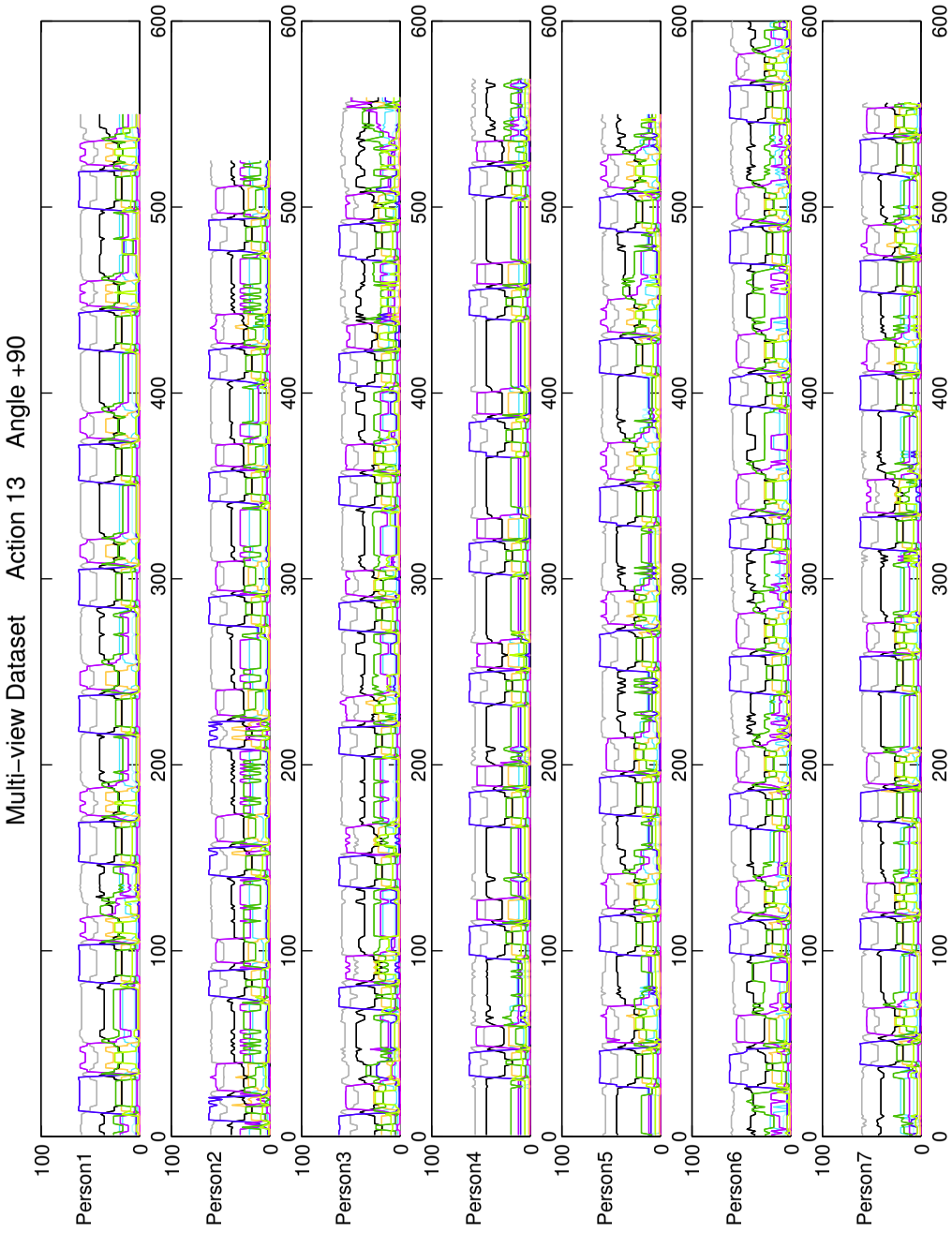

Figure 5.16: Actions from Dataset 2 after clustering

5.3 Recognising actions

There are two general approaches to human action recognition: recognition by reconstruction and direct recognition [54]. The former decomposes the motion into characteristic key frames which implies a sequential pose based approach. Direct recognition tries to recognise the motion directly using features extracted from the collection of image frames comprising the motion.

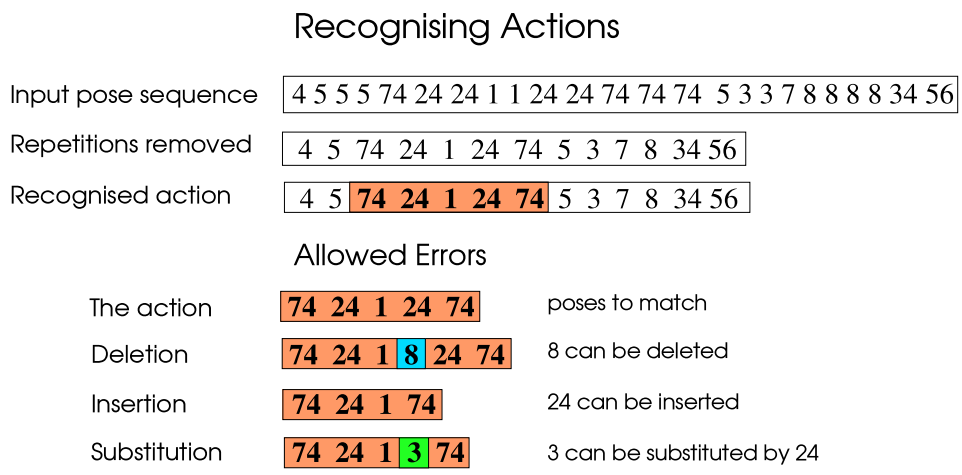

Figure 5.17: Actions are recognised by detecting the action pose sequence in the input stream. One insertion deletion or substitution error is allowed.

Section 5.2 describes how a negative space pose classifier can be constructed. To classify a new image sequence, each image in turn is segmented and pre-processed using the colour-coded representation. The appropriate colour and bounding box features are then extracted and a simple Euclidean distance measure is used to determine the closest pose cluster centre to label the data point. Each image frame in the sequence is now represented simply by its pose label.

As a person performs an action, his or her body goes through a sequence of poses. In the case of a video stream, a number of frames would contain very similar images and thus sequential poses with the same pose label would likely result. By ignoring such sequentially occurring, duplicate pose labels, a compact sequence of poses is obtained. It becomes, for action recognition, merely a task of extracting the unique pose sequence that characterises a specific action. For example, a wave could be characterised by poses

represented by the pose label sequence [74 24 1 24 74], by simply detecting this sequence of numbers in the input stream we have effectively recognised the wave.

An additional benefit of this approach is that the recognition of an action now becomes independent of the duration over which it is performed. Furthermore, there is also no need to find the start and end of an action as is required by some other recognition schemes.

In a captured video stream one is subject to limitations such as the dropping of frames by a frame grabber, a low fame rate or image segmentation errors. In these cases it is possible that a person performs a fast movement that is not captured completely by the camera, or that an individual can perform part of a gesture atypically. Thus the system has to make allowance for the resulting inconsistencies. In the case of the action recognition system, this is compensated for by allowing only one insertion, deletion or substitution error per sequence as illustrated in Figure 5.17.

The tasks of constructing the pose classifier and recognising an action from an input stream have been automated. The system automatically recognises the actions, but the actual sequence of poses that characterise an action group is determined manually by making a visual comparison of the training action sequences. This is possible because the action sequences in the training set contain at most 5 sequential pose class labels meaning that they can easily be identified. However, as the amount of actions increase, this becomes more tedious an exercise and would have to be automated for future practical systems . The system has occasionally labelled more than one sequence per action, as some people have performed a given action too differently. For example, In action 1 three actors had their legs together, while the rest stood with their legs apart resulting in different pose groups. The action is nominally the same, but two different labelled sequences are used to identify these performance variations present in action 1. Table 5.1 contains labelled sequences that characterise the actions contained in Dataset 2.

Action 1a	102	120	46	120	102
Action 1b	50	189	2	189	50
Action 2	62	76	62		
Action 3	193	9	214		
Action 5	250	161	28	161	250
Action 6	395	120	285	120	395
Action 8	167	162	36	162	167
Action 9	205	76	32	76	205
Action 11a	2	123	113	273	
Action11b	165	102	192	21	
Action 13	267	363	40	363	267

Table 5.1: Numbered sequences that characterise actions in Dataset 2.

Once this sequence of labels has been identified for the different actions in Dataset 1 and 2 the system performance can be evaluated. The next section relates what data was used for recognition while Chapter 6 provides the results of the action recognition.

5.3.1 Recognising actions from the datasets

Dataset 1

The test set consisted of 5942 images. The labelled actions of the feedback set makes it simple to establish what the pose sequences for the different actions should be because the feedback set contains samples of all the actions to be recognised in the set. By labelling the testing data and detecting these sequences in the input stream we can recognise the actions. The results will be discussed in the following chapter.

Dataset 2

An additional number of multi-view action sequences are needed to determine the pose sequences that characterise the additional, multi-view actions in this database. For each action, one repetition from 2 different people was removed from the testing set. These actions were then labelled and their pose sequences compared. This was only done for the angle from which a particular action could best be viewed. The number of views that were trained for action recognition were limited because of the overwhelming number of

samples that would have to be evaluated. Testing evaluates how well the system recognises actions from the view it was trained on and how well the system copes with slight differences in view angle. There are 4879 images in the training set and 326733 images in the testing set.

Not all the actions from Dataset 2 were considered for recognition because of the complexity of these actions. The excluded data are Action 4, 7, 10, 11 and 12. Action 10 was excluded as the action is basically composed of a single pose. The actors performed Action 7 in too many varied ways to consider it a single action as seen in the CD sequences. Actions 4, 11, and 12 were left out of the scope of this thesis as they are too complex. Complexity is defined here by the number of times an action "crosses itself". In Action 4, for example, a person starts at the centre position, moves to the left, returns to the same centre position, moves off to the right and then moves back again to the centre pose. Thus the action loops onto itself. Although the poses can be recognised, a more complex action recognition scheme would have to be devised as it is time consuming to determine what the pose label sequence should be. For other actions, the sequence can be determined by knowing what the start and end poses are, the poses in between thus constituting the pose label sequence.

This Chapter explains how the feature data can be partitioned to construct a pose classifier. This classifier is then used to automatically label input sequences according to the poses contained in the images. The actions characterised by these sequences are then automatically recognised by detecting a numbered sequence corresponding to that of a specific action. The following Chapter discusses the results of the action recognition system.

Chapter 6

Results

This chapter relates the results from the negative space pose classifier and action recognition systems. These recognition systems are used to recognise actions from Databases 1 and 2, while the pose classifier is used to direct an animated character and to classify static poses in an image database.

6.1 Recognising actions from Database 1

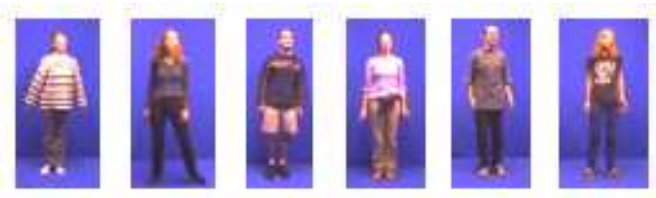

Figure 6.1: The people used in sequences from Database 1.

The negative space colour coding method was used to recognise actions from Database 1 described in Section 3.1. The training set was made up of two women performing free form actions such as waving, lying down and kicking. The only restrictions placed on the types of clothing that could be worn were that dresses were not permitted, and that the colour of the clothing should be different from the background.

The testing set consisted of nearly 6000 image frames compromising 2 male and 2 female students of varying stature and attire. The actions recognised were; waving hands next to the head and waving hands above the head, low and high right and left leg kicks, left and right arm waves and lying down and sitting up, viewed from the left and right hand sides. Figures 4.19, 4.17 and 4.15 show poses from the testing sequences and Appendix D, the accompanying CD, contains animations of the actions sequences. These actions were repeated a number of times by each of the actors.

The pose classifier used 1135 different 11-D feature vectors for training and identified 83 different pose classes. The action classifier for Dataset 1 recognised 14 different classes of actions. Table 6.1 shows that a total of 125 repetitions of the actions were performed and 123 were recognised successfully, giving an accuracy rate of 98%. Two of the actions were not recognised. The first where the right arm is dropped. The action was performed too fast for the camera frame rate, too few poses represent the action and the action pose sequence was not detected. The second action is found in the set of actions where people lie down facing left, the person in question lifted her hand to her face and as a result the image sequence was also not recognised. There were no false recognitions where an actions is detected that is not present in the sequence.

Action	Number Performed	Recognised
Left hand lifted up	9	9
Left hand dropped	9	9
Right hand lifted up	7	7
Right hand dropped	7	6
Waves below the head	9	9
Waves above the head	10	10
Sitting up facing left	12	12
Lying down facing left	12	11
Sitting up facing right	12	12
Lying down facing right	12	12
Low kicks, left side	11	11
High kicks, left side	2	2
Low kicks right side	2	2
High kicks, right side	11	11

TOTAL	Recognitions	123
	Repetitions	125
	Accuracy	98%

Table 6.1: Actions in Database 1

The recognition rates show that negative space processing can achieve good action recognition results on actions parallel to the camera plane. This is especially so given that there were no males in the training set and yet the system had no difficulty in recognising their actions.

The test data set also contained actions not performed in the training set. An example of such an action was where one arm is extended above the head and then dropped to the side. The action itself was not present in the pose classifier training set but the poses that comprise the action were. The new action is correctly classified because the correct pose sequence needed to make up the action was in the training set. Thus, a benefit of the pose approach is that new, untrained actions can be correctly recognised if they are made up of poses present in the classifier. The pose approach also simplifies the action recognition problem as it automatically becomes time independent and obviates the need for a recognition model that has to deal with these variations.

6.2 Recognising actions from Database 2: The Multi-view Sequences

The multi-view dataset (Dataset 2) discussed in section 3.2 contains actions performed by seven people. The actions are captured from six different camera angles. The accompanying CD contains a number of action sequences from Dataset 2. This animated database can be used to study the differences that occur when the people performing the same action are viewed from different view angles. The action recognition system for Database 2 was trained to recognise a selected number of actions. The system is trained to recognise Actions 1, 5, 6 and 13 from the 90 degree angle, and actions 2, 3, 8, and 9 are recognised from the 0 degree, frontal angle. The system was trained on these angles as the actions and poses exhibit the most variation when viewed from this angle. All the other view angles will be less characteristic of the action and it is expected that the recognition rate diminishes the further a testing sequence is from the trained view. These sequences show how well the action recognition system copes with changes in the view angle.

Action		-60	-30	0	+30	+60	+90
1	Recognised	0	0	0	6	49	51
	Repetitions	51	51	51	51	51	51
	Percentage	0	0	0	11.8	96	100
2	Recognised	14	43	52	49	0	0
	Repetitions	52	52	52	52	52	52
	Percentage	26.3	82.7	100	94.2	0	0
3	Recognised	31	36	40	33	13	0
	Repetitions	46	46	46	46	46	46
	Percentage	67.4	78.2	87	71.7	28.2	0
5	Recognised	0	0	0	0	3	45
	Repetitions	53	53	53	53	53	53
	Percentage	0	0	0	0	5.7	85
6	Recognised	0	0	0	0	25	42
	Repetitions	53	53	53	53	53	53
	Percentage	0	0	0	0	47	79
8	Recognised	38	45	48	41	10	3
	Repetitions	52	52	52	52	52	52
	Percentage	54	86.5	92	79	19	5.8
9	Recognised	34	44	45	44	35	0
	Repetitions	50	50	50	50	50	50
	Percentage	68	88	90	88	70	0
13	Recognised	0	0	0	0	3	48
	Repetitions	53	53	53	53	53	53
	Percentage	0	0	0	0	5.7	91

TOTAL for angles trained on	Recognised	371
	Repetitions	410
	Percentage	90.5%

TOTAL for angles 30 degrees from the trained view	Recognised	415
	Repetitions	610
	Percentage	68%

Table 6.2: Recognition results for the multi-view Dataset 2, the red numbers indicate recognition results on view angles that the system is trained to recognise. Green numbers indicate recognition rates for actions 30 degrees removed from the training view. A repetition indicates one complete cycle of an action.

Table 6.2 shows the results from the action classifier. The numbers indicated in red show recognition results for the view angles that the system has been trained on. The total at the bottom of the figure relates an overall recognition result of 90.5% for these angles. The rest of the entries show how well the action is recognised from different view angles.

This result is expected to decrease as the angular distance from the trained view angle is increased and less of the characteristic action variation can be seen. The recognition rate drops off at different rates for an increase in view angle from the trained view. The values in green indicate angles 30 degrees from those that the system has been trained to recognise. The recognition rate on all these untrained angles are 68%. If action 5 and 13 are removed because the probability of recognition decreases most rapidly for these actions the recognition rate increases to 83.5%.

Action		-60	-30	0	+30	+60	+90
1	False	0	0	0	0	0	0
	Repetitions	51	51	51	51	51	51
	Percentage	0	0	0	0	0	0
2	False	0	0	0	0	0	0
	Repetitions	52	52	52	52	52	52
	Percentage	0	0	0	0	0	0
3	False	1	7	4	1	0	0
	Repetitions	46	46	46	46	46	46
	Percentage	2.2	15.2	8.7	2.2	0	0
4	False	0	0	0	0	0	0
	Repetitions	52	52	52	52	52	52
	Percentage	0	0	0	0	0	0
5	False	0	0	0	0	0	0
	Repetitions	53	53	53	53	53	53
	Percentage	0	0	0	0	0	0
6	False	0	0	0	28	16	3
	Repetitions	53	53	53	53	53	53
	Percentage	0	0	0	52.8	30.2	5.7
7	False	1	0	0	0	0	13
	Repetitions	52	52	52	52	52	52
	Percentage	1.9	0	0	0	0	25
8	False	0	0	0	0	0	3
	Repetitions	52	52	52	52	52	52
	Percentage	0	0	0	0	0	5.7
9	False	1	0	2	1	0	0
	Repetitions	50	50	50	50	50	50
	Percentage	2	0	4	2	0	0
11	False	1	8	24	9	0	0
	Repetitions	52	52	52	52	52	52
	Percentage	1.9	15.4	46	17.3	0	0
12	False	0	0	0	0	0	0
	Repetitions	52	52	52	52	52	52
	Percentage	0	0	0	0	0	0
13	False	0	0	0	0	0	0
	Repetitions	53	53	53	53	53	53
	Percentage	0	0	0	0	0	0

TOTAL	False recognitions	123
	Repetitions	3973
	Percentage	3.1%

Table 6.3: False Recognition results from the multi-view Database 2

Table 6.3 shows the false recognition rates. This table contains all of the actions, including those that have not been used for action recognition at all. The action recognition system is then used to see if any of the trained actions are mistakenly recognised in one of these sequences. The overall false recognition rate is 3.06%. The worst recognition as well as false recognition rate is for action 6. This action is always confused with action 1, which is similar.

6.3 Applications

The next section describes two applications using negative space colour coded pose recognition, the directable character Alice and an automatic database pose labelling system.

6.3.1 Pose recognition visualisation using a directable character

The UCT chroma-key room was used to control an interactive directable character. Image frames are captured and sent to an on-line segmentation module that provides the silhouette images used by the pose recognition system. Recognised poses in turn are sent to ALICE [5], a 3D authoring system, which then moves the model limbs into positions mimicking those of the actor. ALICE is a 3D Interactive Graphics Programming Environment built by the Stage 3 Research Group at Carnegie Mellon University. ALICE provides a set of Python classes for dynamically linking the output of the pose recognition system to manipulate characters inside the 3D ALICE environment.

Figure 6.2: The directable characters Alice on the left, and Muis on the right hand side. Alice is a 3D character bundled with the ALICE [5] software, while Muis was constructed specifically for this thesis.

To obtain the directable character, a model is first constructed in a 3D design package such as 3DMAX. The character is then placed into an ALICE world. Alice and Muis are 3-D characters that can be directly rotated and moved in accordance with the recognised pose data. A character has to be designed and hierarchically linked, thus providing a model where the body parts such as the head and different limbs can move independently while still remaining linked to the whole. The output of the pose recognition system consists

of a list of data each corresponding to the overall 3D angle of rotation of a particular limb. This data is then used to move each of the character's limbs to the appropriate positions. Since the parts are hierarchically linked, the command, for the character called Alice, "alice.rightarm.rightforearm.moveto(0,0,0)" will move the Alice's right forearm to point to the origin, relative to the right arm position. By replacing the co-ordinates with variables, the different positions can be read directly from the Matlab pose recognition output file specifying the limb positions. Once these have been obtained the model can be moved automatically in accordance with this data. This procedure is defined for each frame in ALICE and therefore allows the character to move automatically to assume the pose. Interpolation between the frames are automatically done by the ALICE software. The directable character provides direct feedback for the pose recognition system.

6.3.2 Classification on the De Beers database

The pose classification method described in this thesis was used to group similar poses in the De Beers database. This database consists of nearly two thousand images taken in a green Chroma-key studio. These images had to be segmented and then classified into groups based on both the uniform colour and pose information. Manually segmenting and ordering the images would have been time consuming.

Figure 6.3: Poses from the de Beers database

The Colour coded negative space pose clustering method, and adapted Chroma-key segmentation algorithm for a green room made it possible to automate this procedure. The actors were grouped into 7 pose groups: Lying down, upright, crouching or stooping, stepping, left hand extended, right hand extended and poses where a broom is present. This application highlights the complexity of the pose recognition problem. When grouping a dataset into a number of characteristic poses it becomes difficult to assess the success rate. In an action recognition system it is possible to use the labelled actions to evaluate the fitness of a particular partitioning. Although no poses were conspicuously classified incorrectly, it would be too subjective to evaluate the system performance. The very fact that the poses are taken from a dynamic moving entity implies that there can be no definitive pose classification system. When does a stand become a stoop? There are no hard and fast borders between poses so these answers remain elusive, and pose recognition system will have to base its success on whether the subjective needs of the developer have been satisfied.

Chapter 7

Conclusions

7.1 Summary of the contribution

The problem of human action recognition has been addressed in the literature to various degrees of success. What all the methods have in common is that they either investigate the entire image or the space occupied by the body of the person for recognition.

A study of a number of binary silhouette images led to the hypothesis that the poses could more readily be identified from the areas immediately surrounding a person rather than using the space occupied by the body. If a person moves her arm away from her body, the limb merely undergoes a change in the angle of orientation to the body. The spaces surrounding the person, however, show greater changes in more regions as a result of the movement. To date there has been no investigation into the use of these negative spaces in terms of pose and action recognition.

This dissertation addresses the issue of human pose and action recognition using the negative space, particularly the spaces formed between a silhouette and what has been described as its bounding box. Two different Chroma-key environments were constructed to create action sequence databases. The first, Database 1, contains actions captured on site in a Chroma-key studio using only one frontal viewing angle. Database 2 was captured in a film studio, resulting in a dataset consisting of a Chroma-key environment that provided footage from six different camera angles. Eight individuals performed various actions for a set number of repetitions. Both the environments use a Chroma-key background to

simplify the segmentation process. No information relating to the physical space or size of the person was used, implying that any segmented silhouette image of a human could be used as input.

The binary silhouette image is preprocessed using a directional scanning method to form the negative space colour coded image. The colours reveal visually how distinct the negative spaces are for different poses. Features extracted from the negative space representation are based on the percentage area occupied by the colours in the images as well as the bounding box proportions. The effect that an action has on the features can be seen in the feature plots in Chapter 4 and Appendices A and B. The plots illustrate the differences that occur when people perform the same actions. These variations can be seen in the shape of the coloured traces as well as the duration it takes to complete an action. These plots also show that action traces are characterised by the relative amounts of colours present.

People perform actions in varied ways and over different periods of time. These variations make it difficult to recognise an action in its entirety. Actions are composed of a sequence of characteristic, sequential poses. An action can be recognised, irrespective of the variations in duration, if the sequence of poses it moves through can be identified. To label the feature data according to the pose that it represents, a pose classifier needs to be constructed. The design of such a pose classifier is complicated because the feature data is derived from a dynamically changing entity. The feature data itself does not contain natural clusterings and is unlabelled, making it unfeasible to use traditional supervised or unsupervised clustering methods directly to identify the pose classes.

Although we cannot obtain pose labels, it is possible to label data segments according to the actions they contain. Similar action segments are used to provide feedback for the pose clustering algorithm, by comparing how similar the labelled feedback sequences are for the same action. This similarity measure is based on a longest common sequence measure and a measure that ensures that actions are not described by too few poses. Different partitions will result in different sequence labels. Too few pose classes will fail to adequately describe the action, while too many could result in a different pose label for each data point. With too many clusters even very similar actions will be described by a different sequence of pose labels. The best partitioning is the one that results in the same

labelled pose sequences for the same action.

The pose partitioning did not result from natural clusters and it is therefore expected that a number of pose partitions might have neighbouring regions containing data derived from very similar images. These regions need to be merged to avoid a classifier that produces distinct labels for poses that are virtually indistinguishable. Information independent of the classifier is used to merge clusters if they contain similar poses. Each cluster can be represented by a cluster centre, and this cluster centre can be associated with a negative space coloured image from the nearest data point. Neighbouring clusters can be merged if the images representing the clusters, are highly correlated. By again using the feedback action set, it is possible to find the best correlation threshold for merging clusters. By using a feedback action set to evaluate the classifier performance, it becomes possible to design a system that is very difficult to construct using only traditional clustering methods.

A new sequence can now by classified according to the poses it contains by extracting the negative space features from the image and assigning them to the nearest cluster centre. Each cluster centre is assigned a number, and thus the image sequence can be represented by a numbered sequence. The action classifier is constructed using the output from the pose classifier. To construct a classifier for an action, a number of sequences of different people performing the same action are labelled using the pose classifier. These sequences should have very similar pose labels as they describe the same action. The string of poses that most generally represents the action is identified. This is repeated for all the actions to be recognised. The action classifier thus consists of a number of strings that represent the actions. It monitors the new incoming poses to see if the sequence matches that of one of the actions. By recognising this unique string of poses the system is able to recognise the action performed in the footage.

The system has been tested on action sequences from the front view Database 1 and multi-view Database 2. The front view dataset was used to determine the sensitivity of the system to changes in personal appearance. The multi-view dataset investigated how well the system recognised actions from view angles different to those that the system was trained to recognise. The only constraint placed on the appearance of the actors was that they should not wear dresses or blue, the colour of the background.

The front view dataset used a training set consisting of two women performing free form actions to construct the pose classifier. The pose classifier distinguished between 83 different pose groups. This was used to classify the actions of two men and differently attired women. There were 14 different actions each performed a number of times. In total, 123 out of 125 actions were recognised, a recognition rate of 98% with no misclassified actions. This demonstrates that the negative space features are not sensitive to differences in personal appearance.

The multi-view dataset consisted of seven individuals, three women and four men. The actions are performed several times and captured from six different camera angles each separated by 30 degrees. The pose classifier categorises 195 different multi-view poses. These poses are used to recognise 8 different actions. For angles from which the action can clearly be seen a total of 371 out of 410 sequences were recognised, a recognition rate of 90.5%. It is difficult to determine how well a system generalises for changes in the view angle, as some actions undergo almost no change for a small difference in view angle whereas others are likely to become unrecognisable. The action recognition system was used to recognise actions for sequences captured 30 degrees from the view that the system was trained on. The system correctly recognised 415 of the 610 possible sequences, a recognition rate of 68%. This rate increases to 83.5% when actions that change too much over the 30 degrees are not considered. A total of 123 sequences were misclassified when the entire database of 3973 sequences were tested, this gives a false recognition rate of 3.1%.

The recognition results on the frontal and multi-view datasets show that poses and actions can be successfully recognised using features extracted from the negative space. The recognition system can cope with small changes in the view angle and differences in personal appearance.

The pose classifier has been used in a number of applications. These applications highlight the difficulties in evaluating how well the classifier performs. One way to determine this is by using the output of the classifier to drive an animated character. A simple directable character was implemented using the ALICE [5] graphics software. This alerts the operator immediately when the pose classifier has failed as the animated character will perform a different pose.

Another application of the pose classifier is to sort an image database into pose groups. This application illustrated how difficult it is to evaluate the performance of a pose classifier. A pose recognition system will have to base its success on whether the subjective needs of the developer have been satisfied.

The work done for this dissertation has shown that the features derived from negative space regions surrounding a person, represents the underlying characteristics of the human body adequately and can be successfully used for pose and action recognition.

7.2 Recommendations and Future work

There are many ways to extend the work done in this thesis. A number of assumptions and constraints have been imposed on the system to investigate if the negative space can be used for pose and action recognition. Improvements and further avenues of research are discussed in terms of the framework of the chapters they appear in.

Chapter 3 describes the image sequence data. The actions in Dataset 2 were unrehearsed, introducing more variation into the data than necessary. This makes it difficult to determine whether the action recognition system has failed or if an action cannot be considered part of that set due to the way in which it is performed. The current system uses Chroma-key segmentation to simplify the extraction of the human silhouette. The problem of segmentation would have consumed too much of the research time. Now that it has been demonstrated that poses and actions can be recognised using negative space , this segmentation method can be improved to allow for recognition in cluttered environments. The current system uses no information relating to physical dimensions of the space or person. This was excluded to ensure that the recognition relied solely on the negative spaces and was not biased by information other than that of the silhouette and bounding box. Features involving the direction and speed of, for example, the bounding box could be extracted to increase the amount of information relating to an action.

Chapter 4 : The work done in this thesis assumes that the camera orientation is at right angles to the gravity, and that the poses arise form a body under the influence of gravity. This assumption ensures the patterns formed by the colours are common to all persons similarly attired in a similar pose. The negative space features are very simple to extract and provide a good representation of the body pose. These features are a very inexpensive way to generate data for other recognition systems.

Chapter 5: The pose classifier was constructed without experimenting with different feature weightings, perhaps such a weighting could result in a better classifier. The clustering method used in this section was a simple K-means method using the Euclidean distance measure. Alternatives to the Euclidean distance measure could also be investigated.

Other clustering methods such as hierarchical and fuzzy clustering can be experimented with. The effect of different weightings on the feature data and clustering can be investi-

gated.

Although recognition has been automated, the extraction of the pose sequences characteristic of an action is still done manually. This should be automated by comparing a number of the labelled action sequences and finding the most common sequences of poses.

It must be remembered that, as stated in Chapter 3 the action sequences considered in this study were derived from controlled environments (the Chroma-key rooms). These have avoided the very important problems a real world implementaion of this technology would face: the preliminary segmentation of the human silhouette which may well introduce variance that would imply an error propagation into negative space area extraction. To deal with this, more complicated action recognition method might be required. Many action recognition systems use a Hidden Markov Model to evaluate the probability of a certain action having occurred, the application of negative space features toward such a recognition system can be investigated to help resolve this. This approach can be used to initialise the pose for more complicated methods that attempt to find detail but needs to have a overall starting pose.

Only one route was taken to investigate the negative space there could be potentially many other avenues and applications. Hopefully the negative space will lead to many interesting investigations.

Bibliography

[1] J. Aggarwal and Q. Cai. Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):428–440, March 1999.

[2] E. Alhoniemi, J. Himberg, J. Parhankangas, and J. Vesanto. SOM Toolbox. http://www.cis.hut.fi/project/somtoolbox.html, June 2000.

[3] J. Amat, M. Casals, and M. Frigola. Stereoscopic system for human body tracking in natural scenes. In *International Workshop on Modeling People at ICCV'99*, Corfu, Greece, September 1999.

[4] R. Arnheim. *Art and Visual perception, the psychology of the creative eye*. University of California Press, 1974.

[5] ALICE Stage 3 Research Group at Carnegie Mellon University. Alice model paint animate 3d graphics for the www. http://www.alice.org, 1999.

[6] A. Baumberg and D. Hogg. Learning flexible models from image sequences. *Lecture Notes in Computer Science*, 800:299–308, 1994.

[7] A. Bharatkumar, K. Daigle, M. Pandy, Q. Cai, and J. Aggarwal. Lower limb kinematics of human walking with the medial axis transformation. In *IEEE Workshop on Non-Rigid Motion*, pages 70–76, 1994.

[8] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE PAMI*, 23(3):257–267, March 2001.

[9] A. Bobick and A. Wilson. A state-based approach to the representation and recognition of gesture. In *IEEE PAMI*, volume 19, pages 1325–1338, 1997.

[10] C. Bregler. Learning and recognizing human dynamics in video sequences. In *Proc. IEEE Conf. Comp. Vision and Pattern Recognition*, pages 568–574, San Juan, Puerto Rico, June 1997.

[11] Q. Cai, A. Mitiche, and J. K. Aggarwal. Tracking human motion in an indoor environment. In *ICIP*, pages 215–218, 1995.

[12] L. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. Technical Report 309, M.I.T Media Laboratory Perceptual Computing Section, 1995.

[13] S. Carlsson and J. Sullivan. Action recognition by shape matching to key frames. In *Workshop on Models versus Exemplars in Computer Vision*, Kauai, Hawaii, December 2001.

[14] Gestalt Chalice. Perception (psychology). http://encarta.msn.com, 2001.

[15] Z. Chen and H. J. Lee. Knowledge-guided visual perception of 3d human gait from a single image sequence. *IEEE Trans. on Systems, Man, and Cybernetics*, 22(2):336–342, March 1991.

[16] German K. M. Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture.

[17] C. Christensen and S. Corneliussen. Visualization of human motion using model-based vision. Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, 1997.

[18] T. Darrel, P. Maes, B. Blumberg, and A. P. Pentland. A novel environment for situated vision and behaviour. In *Workshop for Visual Behaviours at CVPR-94*, 1994.

[19] D.L. Davies and D.W. Bouldin. A cluster separation measure. *IEEE PAMI*, 1(2):224–227, 1979.

[20] R. Duda and P. Hart. *Pattern Classification-2nd ed.* Wiley-Interscience Publication, 2001.

[21] M. C. Escher. Day and night,1938. woodcut in black and gray, printed from two blocks. 39.1 x 67.7 cm. http://www.artchive.com, 2001.

[22] D. M. Gavrila and L. Davis. 3d model based tracking of humans in action: a multi-view approach. In *CVPR*, pages 73–80, San Francisco, USA, 1996.

[23] D. M. Gavrilla. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, January 1999.

[24] L. Goncalves, E. Di Bernardo, and P. Perona. Reach out and touch space (motion learning). In *Third International Conference on Automatic Face and Gesture Recognition*, pages 234–239, Nara, Japan, April 1998.

[25] Haisong Gu, Yoshiaki Shirai, and Minoru Asada. Mdl-based spatio temporal segmentation from motion in a long image sequence. In *CVPR*, pages 448–453, 1994.

[26] J. Gu, T. Chang, I. Mak, S. Gopalsamy, H.C. Shen, and M.M.F. Yuen. A 3d reconstruction system for human body modeling. In *CAPTECH98*, pages 229–241, 1998.

[27] I. Haritaoglu, D. Harwood, and L. Davis. Ghost: A human body part labeling system using silhouettes. In *Fourteenth International Conference on Pattern Recognition*, pages 77–82, 1998.

[28] I. Haritaoglu, D. Harwood, and L. Davis. W4: Who? When? Where? What? : A real time system for detecting and tracking people. In *The Third Face and Gesture Recognition Conference*, pages 222–227, 1998.

[29] B. Heisele and C. Whler. Motion-based recognition of pedestrians. In *International Conference on Pattern Recognition*, pages 1325–1330, 1998.

[30] A. Hilton and T. Gentils. Popup people: Capturing human models to populate virtual worlds. In *Siggraph*, 1998.

[31] D.C. Hogg. *Interpreting Images of a Known Moving Object*. PhD thesis, University of Sussex, UK, 1984.

[32] E. A. Hunter, P. H. Kelly, and R. C. Jain. Estimation of articulated motion using kinematically constrained mixture densities. In *IEEE Non-Rigid and Articulated Motion Workshop*, pages 10–17, Puerto Rico, USA, 1997.

[33] S. Ioffe and D. A. Forsyth. Finding people by sampling. In *Int. Conf. Computer Vision*, pages 1092–1097, 1999.

[34] Y. Iwai, K. Ogaki, and M. Yachida. Posture estimation using structure and motion models. In *International Conference on Computer Vision*, Corfu, Greece, September 1999.

[35] S. Iwasawa. Real-time estimation of human body posture from monocular thermal images. In *CVPR*, pages 15–20, 1997.

[36] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[37] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[38] M. Janse van Vuuren and G. de Jager. Human pose and action recognition. In *PRASA*, pages 29–34, November 2000.

[39] M. Janse van Vuuren and G. de Jager. Art and image processing. In *PRASA*, pages 23–28, November 2001.

[40] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14(2):201–211, 1973.

[41] S. Ju. Human motion estimation and recognition. Technical report, University of Toronto, 1996.

[42] I. Kakadiaris and D. Metaxas. Vision-based animation of digital humans. In *Conference on Computer Animation*, pages 144–152, 1998.

[43] Y. Kameda and M. Minoh. A human motion estimation method using 3-successive video frames. In *International Conference on Virtual Systems and Multimedia*, pages 135–140, 1996.

[44] F. Lerasle, G. Rives, and M. Dhome. Human body limbs tracking by multi-ocular vision. In *Scandinavian Conference on Image Analysis*, Lappeenranta, Finland, 1997.

[45] M. K. Leung and Y. H. Yang. First sight: A human body outline labeling system. *IEEE Trans. on PAMI*, 17(4):359–377, 1995.

[46] M. K. Leung and Y.H. Yang. A region based approach for human body motion analysis. *Pattern Recognition*, 20(3):321–329, 1987.

[47] M.K. Leung and Y.H. Yang. Human body motion segmentation in a complex scene. *Pattern Recognition*, 20(1):55–64, 1987.

[48] G. Loy, J. Sullivan, and S. Carlsson. Pose-based clustering in action sequences. In *IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis*, pages 66–72, 2003.

[49] O. Masoud and N. Papanikolopoulos. A method for human action recognition. *Image and Vision Computing*, 21(8):729–743, August 2003.

[50] S. J. McKenna, S. Jabri, Z. Duric, and H. Wechsler. Tracking interacting people. In *4th Int. Conf. on Automatic Face and Gesture Recognition*, pages 348–353, 2000.

[51] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Trans. P.A.M.I*, 15(6):580–591, June 1993.

[52] A. Mittal, L. Zhao, and L.Davis. Human body pose estimation by shape analysis of silhouettes. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, Orlando, Florida, July 2003.

[53] T. Moeslund and E. Granum. Multiple cues used in model-based human motion capture. In *The fourth International Conference on Automatic Face and Gesture Recognition*, page 362, Grenoble, France,, March 2000.

[54] T. B. Moeslund and E. Granum. A survey of computer-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[55] Saied Moezzi, Arun Katkere, Don Y. Kuramura, and Ramesh Jain. Reality modeling and visualization from multiple video sequences. In *IEEE Computer Graphics and Applications*, volume 16, pages 58–63, November 1996.

[56] Mondrian. Gray tree, 1911. oil on canvas, 78.5 x 107.5 cm. Haags Gemeentemuseum, The Hague. http://www.artchive.com, 2001.

[57] O. Munkelt, C. Ridder, D. Hansel, and W. Hafner. A model driven 3d image interpretation system applied to person detection in video images. In *International Conference on Pattern Recognition*, volume 1, pages 70–73, 1998.

[58] A. Nakazawa, H. Kato, and S. Inokuchi. Human tracking using distributed video systems. In *International Conference on Pattern Recognition*, 1998.

[59] P. J. Narayanan, Peter W. Rander, and Takeo Kanade. Constructing virtual worlds using dense stereo. In *IEEE International Conference on Computer Vision*, pages 3–10, Bombay, India, January 1998.

[60] D. Ormoneit, H. Sidenbladh, M. J. Black, T. Hastie, and D.J. Fleet. Learning and tracking human motion using functional analysis. To appear in IEEE Workshop on Human Modelling, Analysis and Synthesis, June 2000.

[61] J. O'Rourke and N. I. Badler. Model-based image analysis of human motion using constraint propogation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-2(6):523–536, November 1980.

[62] A. Pentland. Smart rooms. *Scientific American*, 274(4):54–62, 1996.

[63] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(7):730–742, July 1991.

[64] F. Perales and J. Torres. A system for human motion matching between synthetic and real images based on a biomechanic graphical model. In *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 83–88, Austin, 1994.

[65] R. Polana and R. Nelson. Low level recognition of human motion. In *IEEE Workshop on Motion of Non-rigid and Articulated Objects*, Austin, TX, 1994.

[66] K. Rohr. Towards model-based recognition of human movements in image sequences. *CGVIP: Image Understanding*, 59(1):94–115, January 1994.

[67] K. Rohr. *Human movement analysis based on explicit motion models*, volume 9 of *Computational Imaging and Vision Series*. Kluwer academic publishers, London, 1997.

[68] R.Polana and R. Nelson. Detection and recognition of periodic, nonrigid motion. *International Journal of Computer Vision*, 23(3):261–282, June 1997.

[69] Gregory Shakhnarovich, Paul Viola, and Trevor Darrell. Fast pose estimation with parameter-sensitive hashing. pages 750–757, Nice, France, 2003.

[70] L. Sigal, M. Isard, B. Sigelman, and M. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[71] L Sigal, Michael Isard, Benjamin H. Sigelman, and Michael J. Black. Attractive people: Assembling loose-limbed. In *NIPS*, 2003.

[72] Marius-Calin Silaghi, Ralf Plänkers, Ronan Boulic, Pascal Fua, and Daniel Thalmann. Local and global skeleton fitting techniques for optical motion capture. *Lecture Notes in Computer Science*, 1537:26, 1998.

[73] A.R. Smith and J.F. Blinn. Blue screen matting. In *SIGGRAPH*, pages 259–268, August 1996.

[74] J.T. Tou and R.C. Gonzalez. *Pattern Recognition Principles*. Addison-Wesley, Massachusetts, 1974.

[75] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *ICCV*, pages 50–59, 2001.

[76] T. Tsukiyama and Y. Shirai. Detection of the movements of persons from a sparse sequence of tv images. *Pattern Recognition*, 18(3/4):207–13, 1985.

[77] R. H. Turi and S. Ray. An application of clustering in colour image segmentation. In *ICARCV*, Singapore, 2000.

[78] A. D. Wilson, A. F. Bobick, and J. Cassell. Recovering the temporal structure of natural gesture. Technical Report 388, M.I.T., 1996.

[79] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. PAMI*, 19(7):780–785, July 1997.

[80] C. Wren and A. Pentland. Dynamic models of human motion. In *FG'98,*, Nara, Japan, April 1998.

[81] C.H. Wren. *Understanding Expressive Action*. PhD thesis, M.I.T., March 2000.

[82] M. Yamada, K. Ebihara, and J. Ohya. A new robust real-time method for extracting human silhouettes from color images. In *Third International Conf. on Automatic Face and Gesture Recognition*, pages 528–533, Nara, Japan, 1998.

[83] M. Yamamoto and K. Koshikawa. Human motion analysis based on a robot arm model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 664–665, Hawaii, June 1991.

[84] L. Zelnik-Manor and M. Irani. Event-based video analysis. In *CVPR*, December 2001.

[85] L. Zhao and C. Thorpe. Recursive context reasoning for human detection and parts identification. In *IEEE HuMAnS*, June 2000.

[86] I. Y. Zheng and S. Suezaki. A model based approach in extracting and generating human motion. In *International Conference on Pattern Recognition*, pages 1201–1205, 1998.

Appendix A

Front view feature plots

A.1 Front view data plots from Dataset

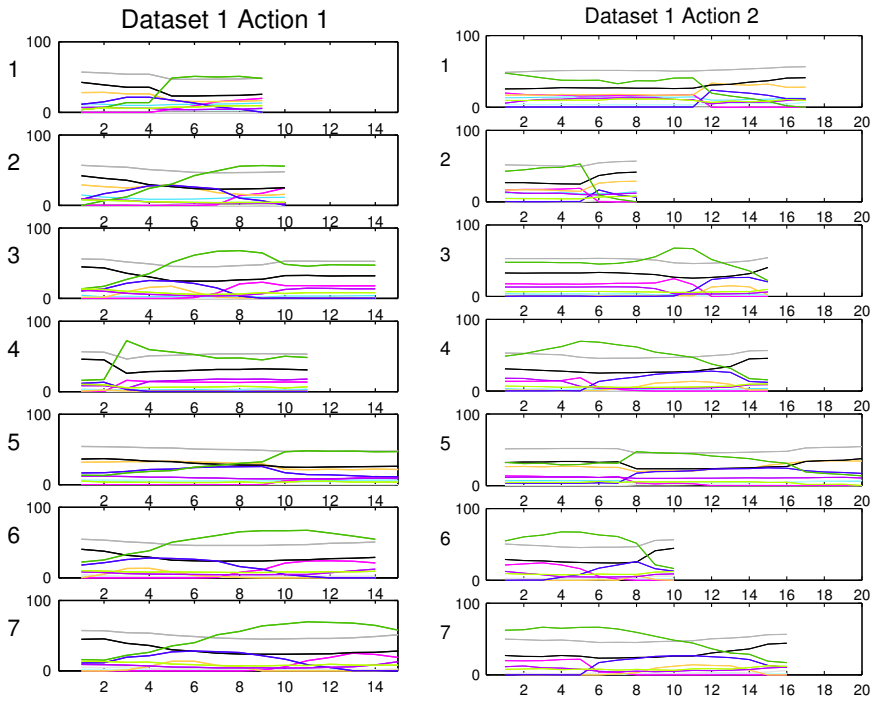

Figure A.1: The traces on the left shows a series of people lifting their right arm up wheres the traces on the right shows it being lowered again.

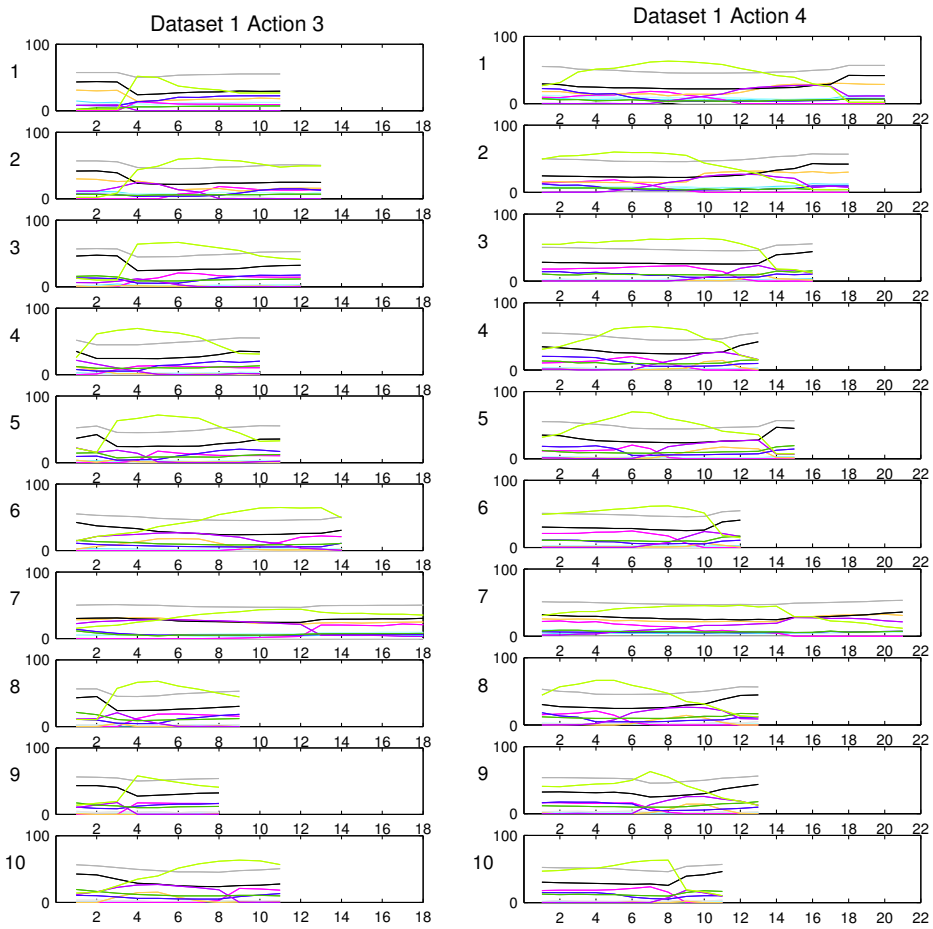

Figure A.2: The traces on the left shows a series of people lifting their left arm up wheres the traces on the right shows it being lowered again.

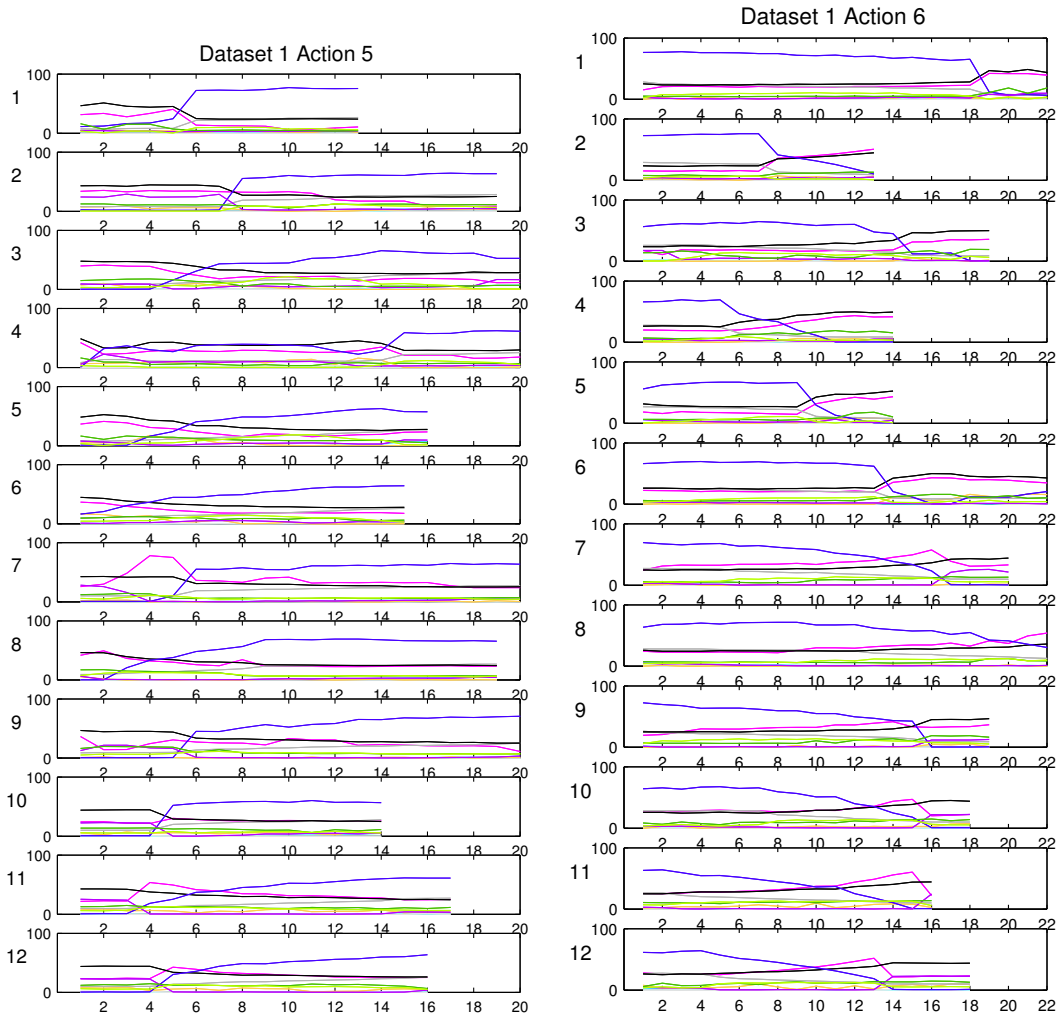

Figure A.3: Traces from Dataset 1 showing people sitting up and lying down facing the left hand side of the image. This action is characterised by an increasing or decreasing dominant blue and black.

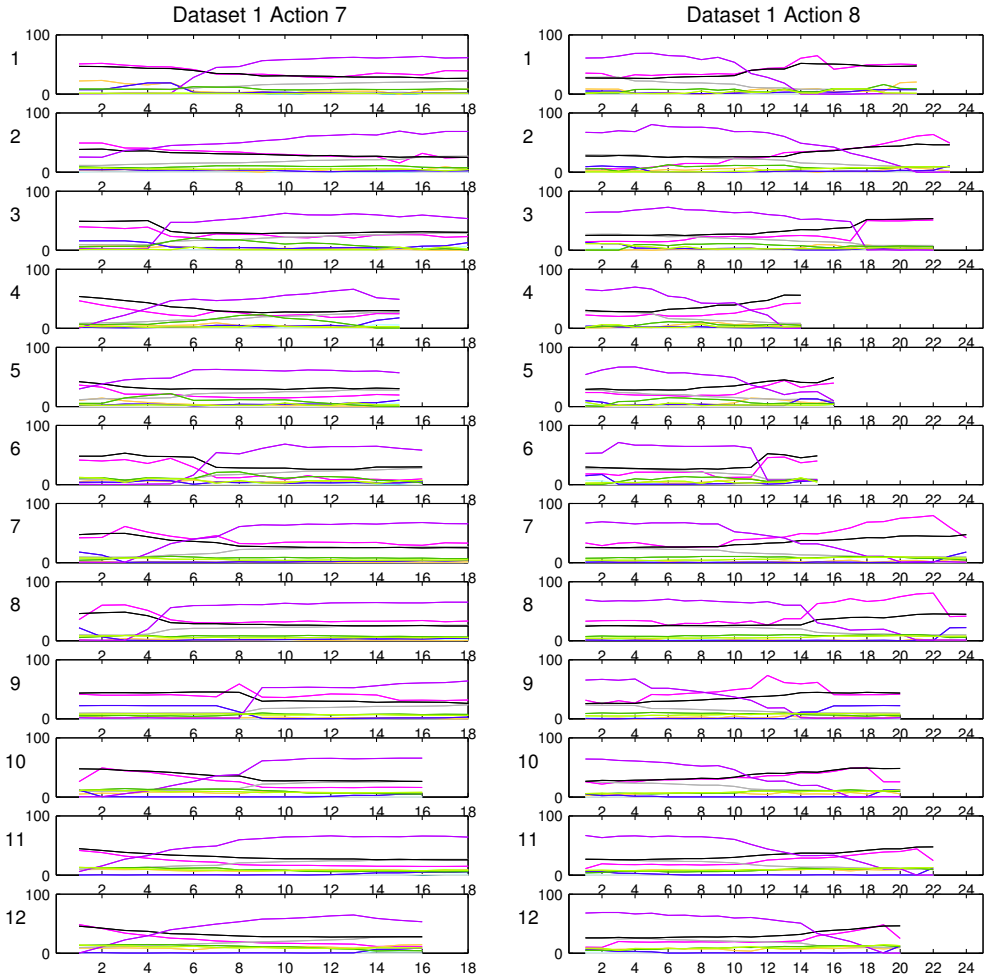

Figure A.4: Traces from Dataset 1 showing people sitting up and lying down facing the right hand side of the image. This action is characterised by an increasing or decreasing dominant purple-blue and black. Traces showing variation within the same action group.

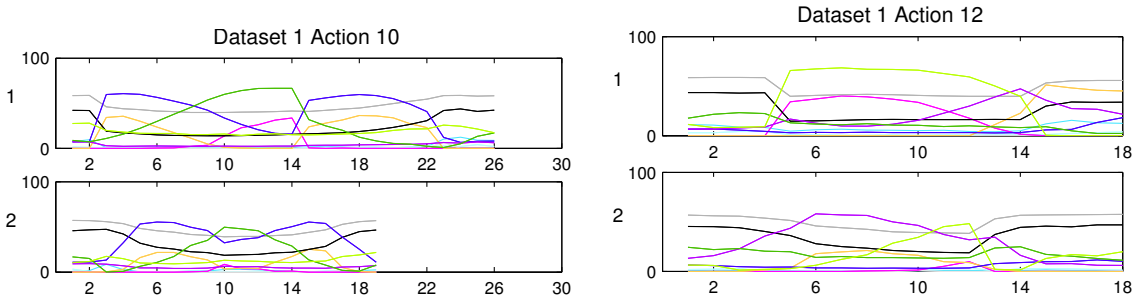

Figure A.5: Action 10 consists of two traces where people perform a very high kick to the right of the body. Action12 represents a high kick to the left.

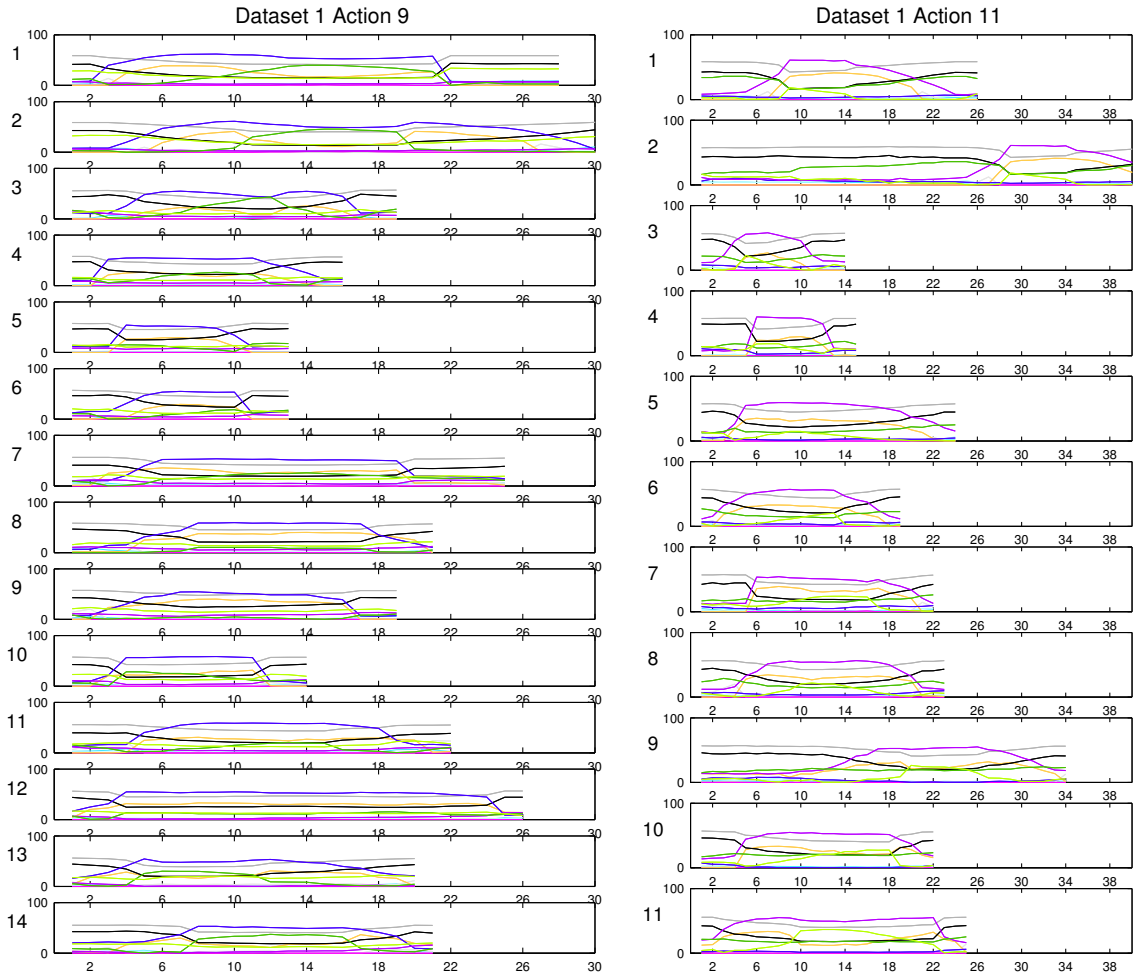

Figure A.6: Traces showing right and left hand side kicks. They can be identified by the large amounts of yellow and either violet or maroon.

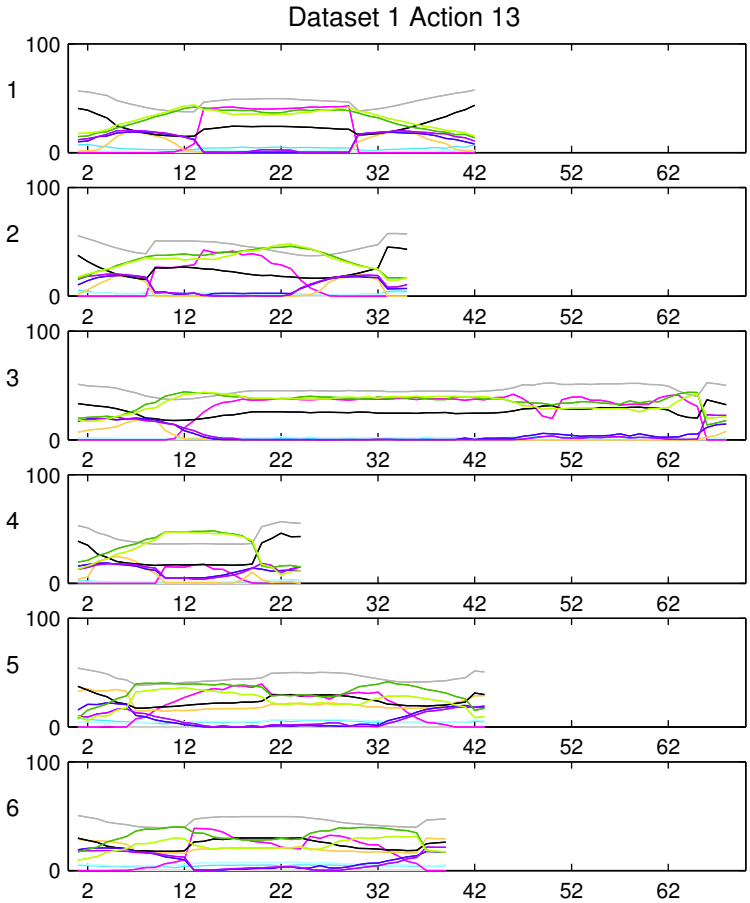

Figure A.7: The actions shows people performing a half wave that turns at the sides of the head. Not the equal amounts of lime and olive and increasing amounts of red toward the centre of the action.

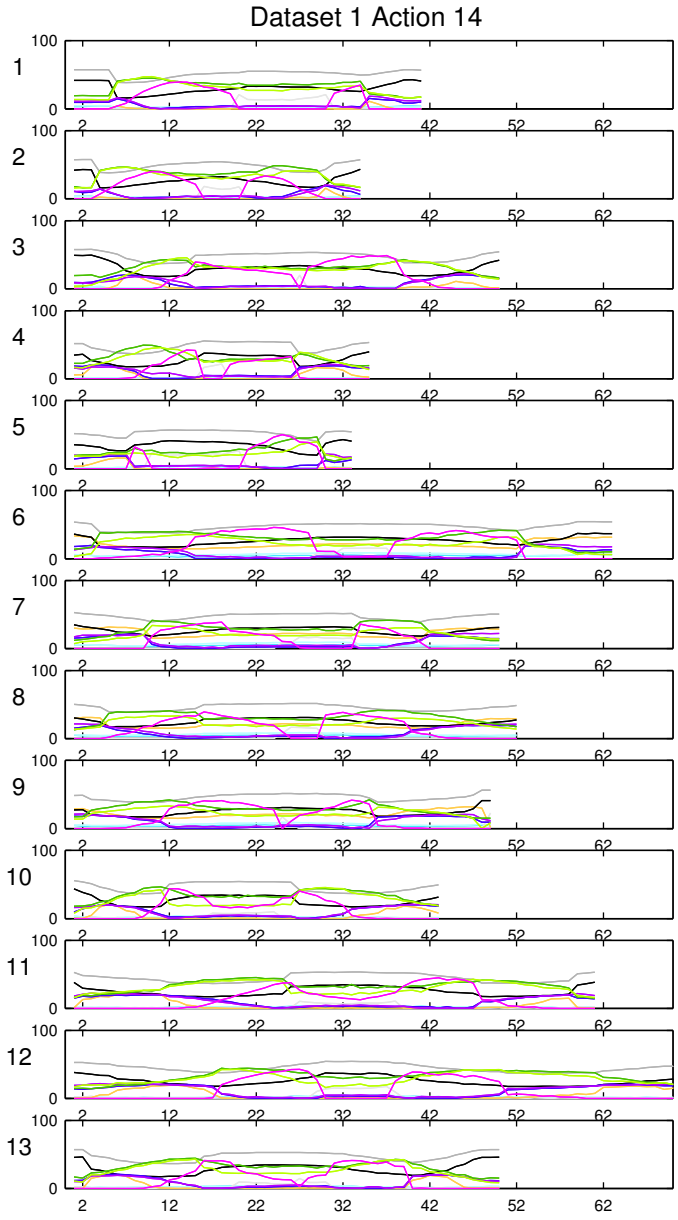

Figure A.8: Traces from Dataset 1 showing a set of full waves. The characteristic red bumps are formed when the arms are above the shoulders and decrease when they are above the head.

A.2 Front view data plots from Dataset 2

The frontview datasets are too complex too explain in words. Appendix D shows animated images of all the sequence actions.

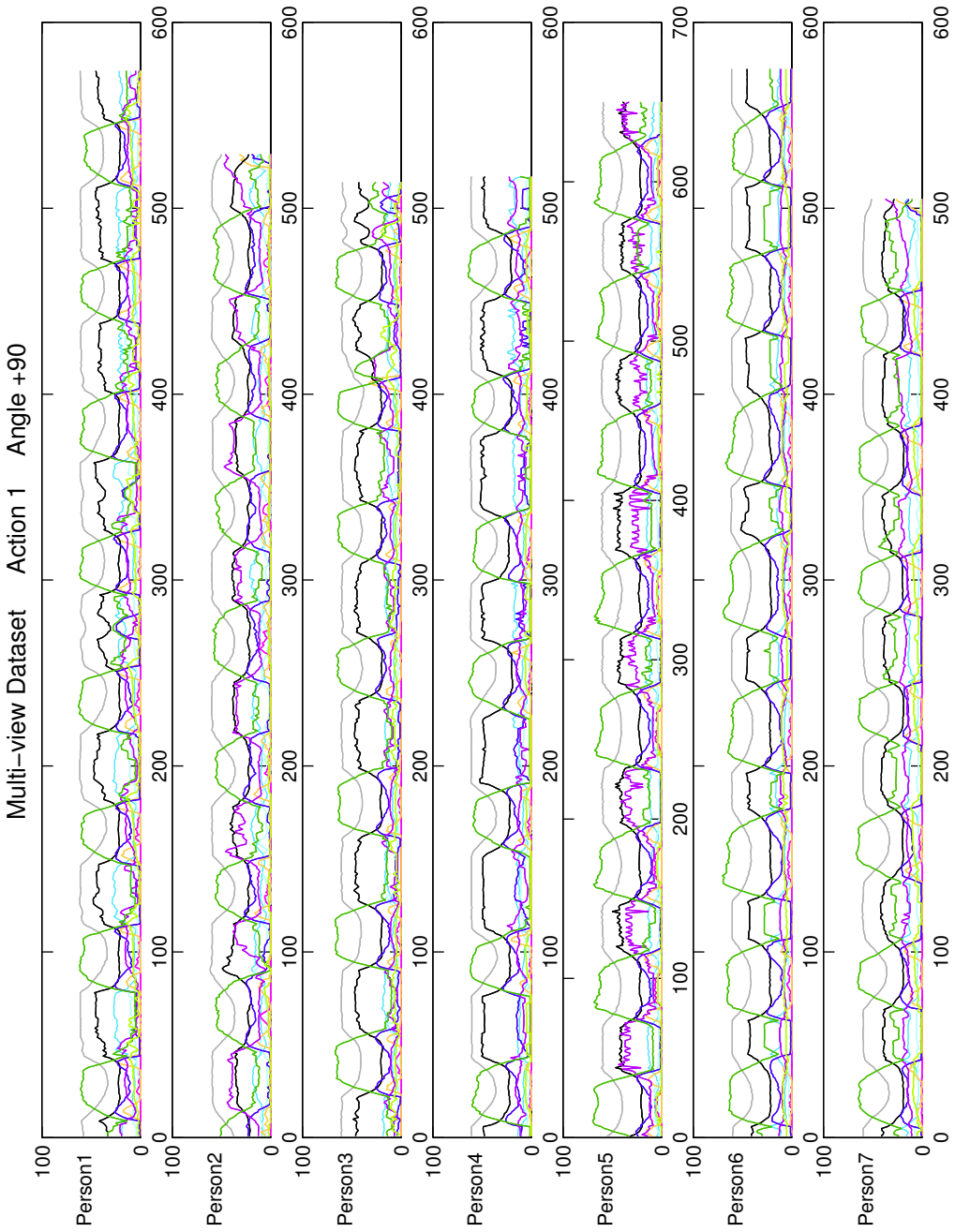

Figure A.9: Feature traces from Action 1 viewed at an angle of 90 degrees showing plots for 7 people performing a number of repetitions of the same action.

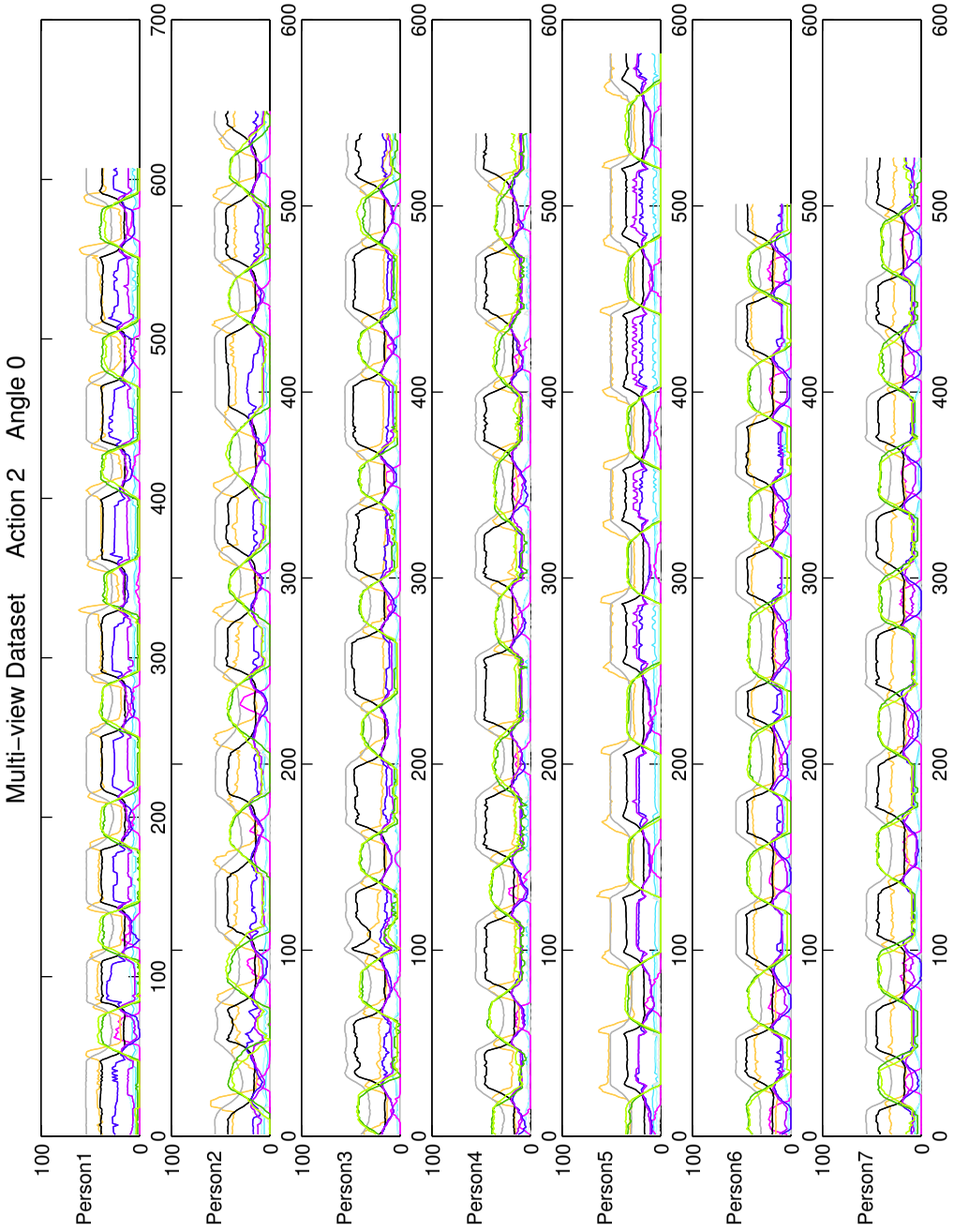

Figure A.10: Feature traces from Action 2 viewed at an angle of 0 degrees showing plots for 7 people performing a number of repetitions of the same action.

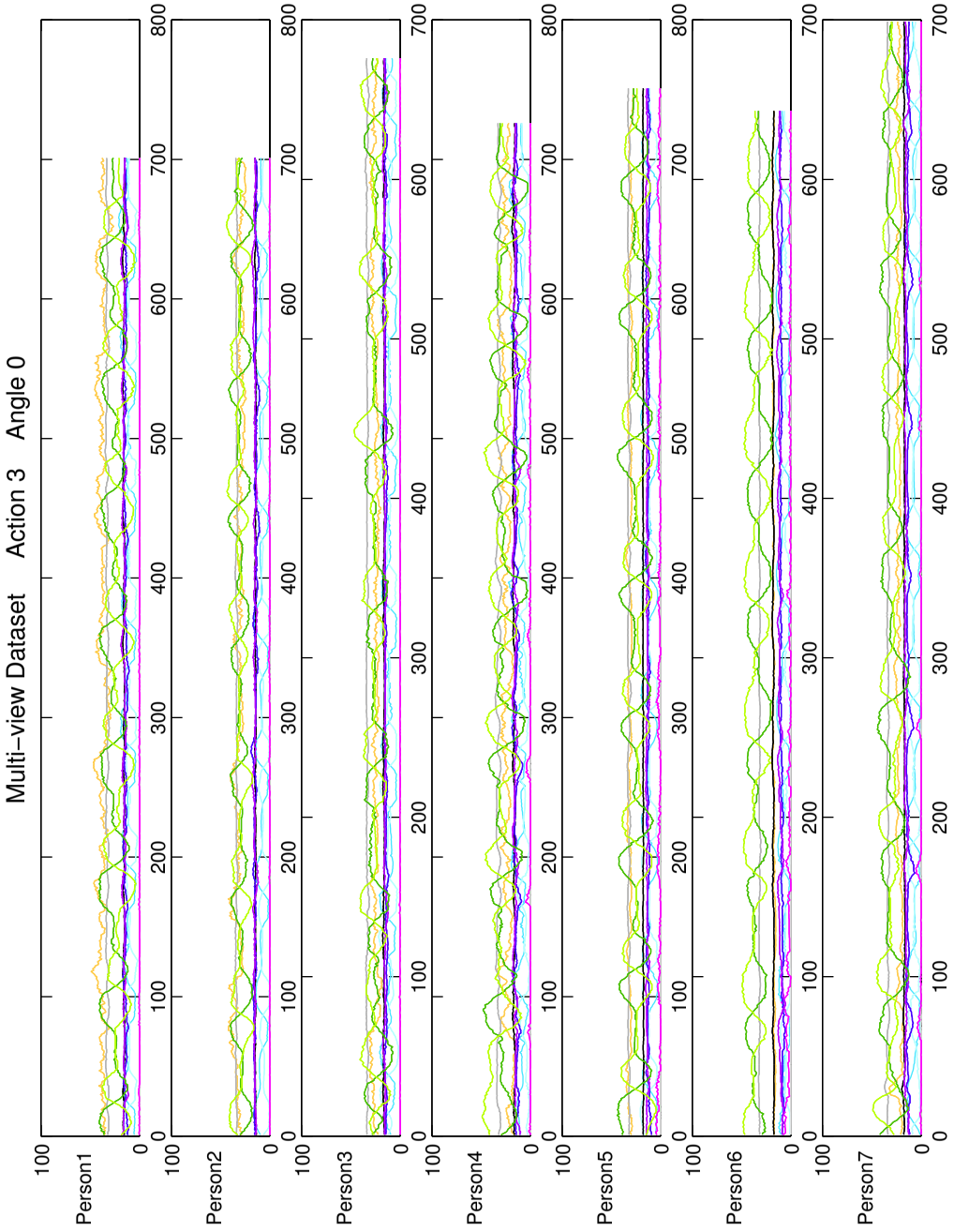

Figure A.11: Feature traces from Action 3 viewed at an angle of 0 degrees showing plots for 7 people performing a number of repetitions of the same action.

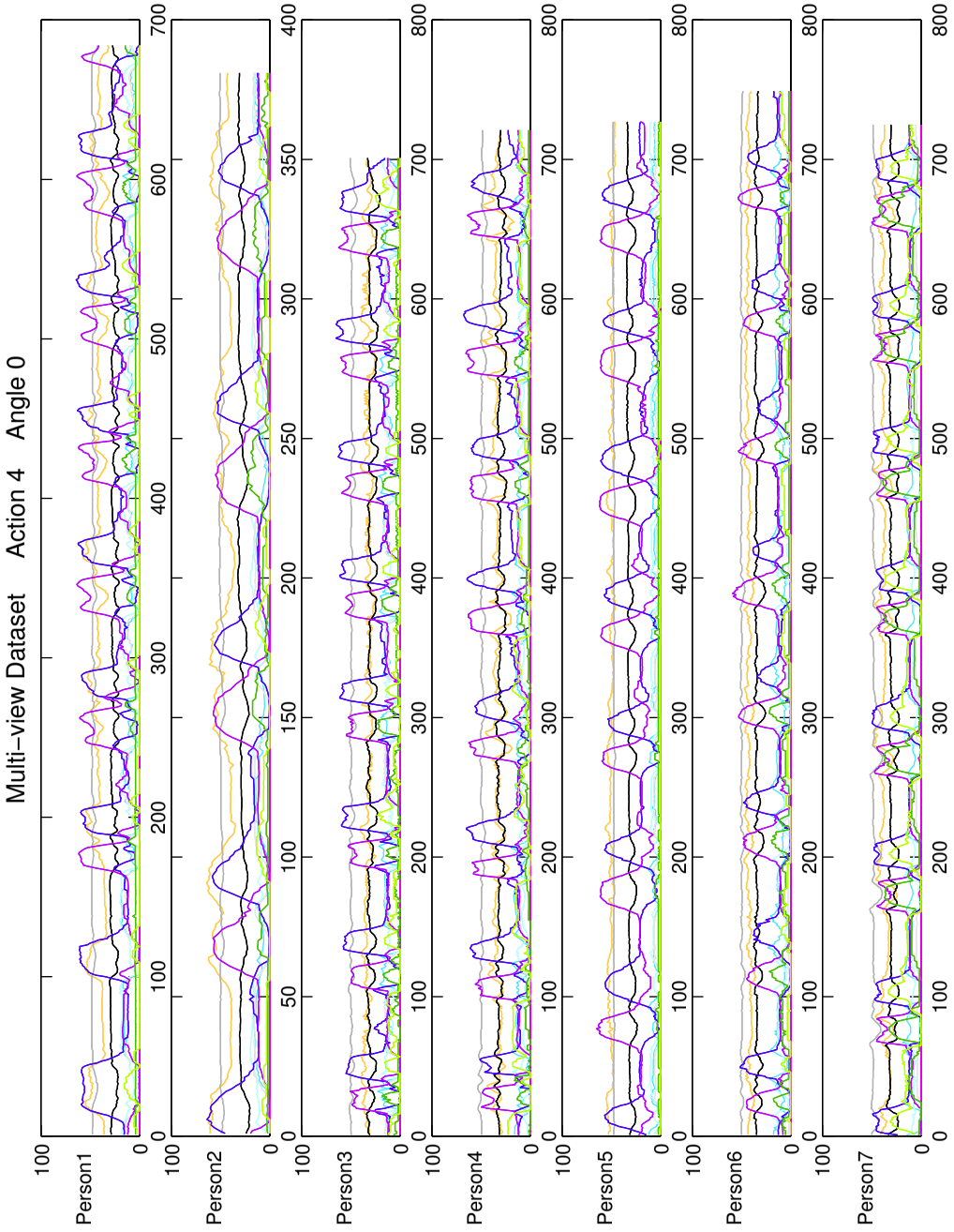

Figure A.12: Feature traces from Action 4 viewed at an angle of 0 degrees showing plots for 7 people performing a number of repetitions of the same action.

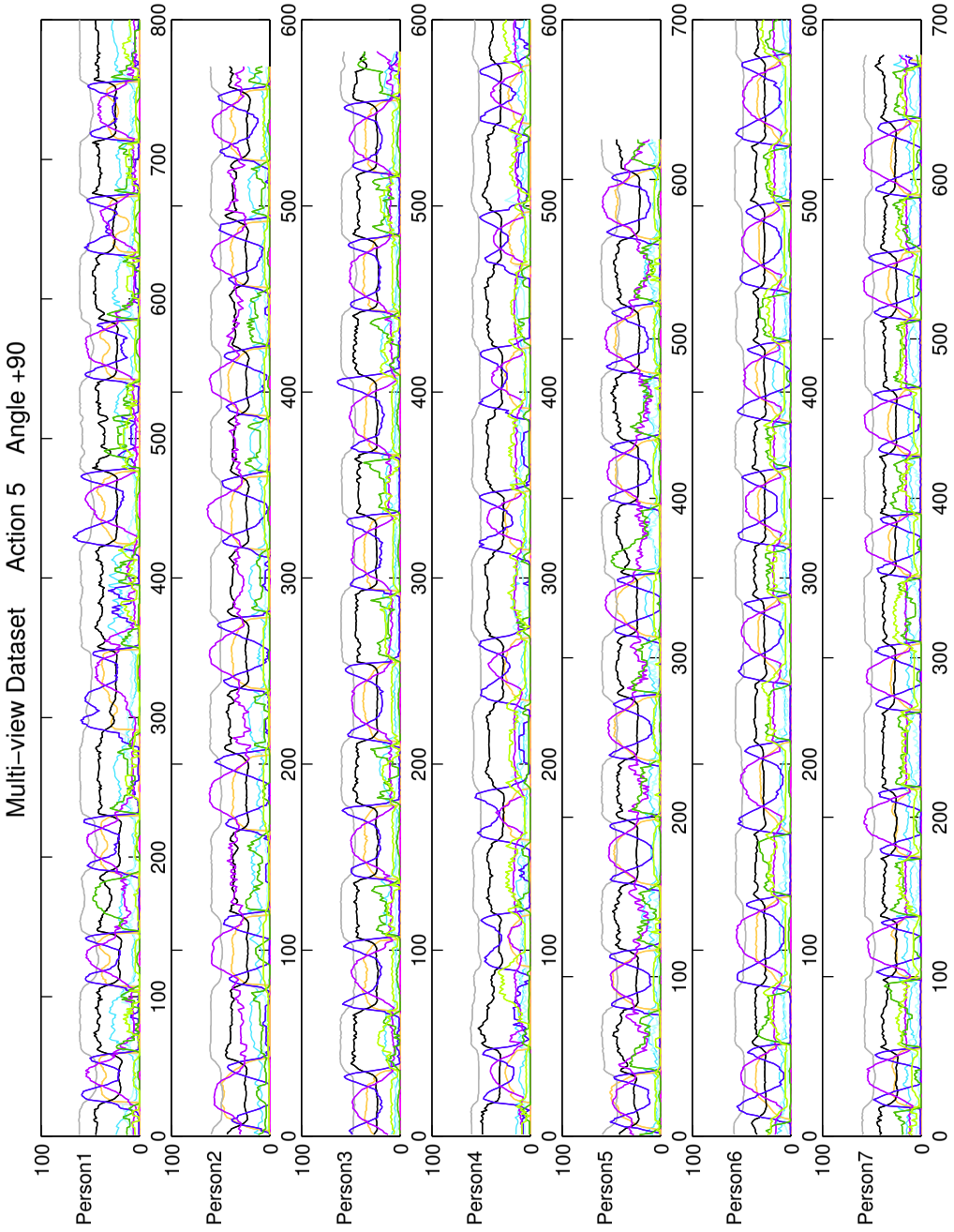

Figure A.13: Feature traces from Action 5 viewed at an angle of 90 degrees showing plots for 7 people performing a number of repetitions of the same action.

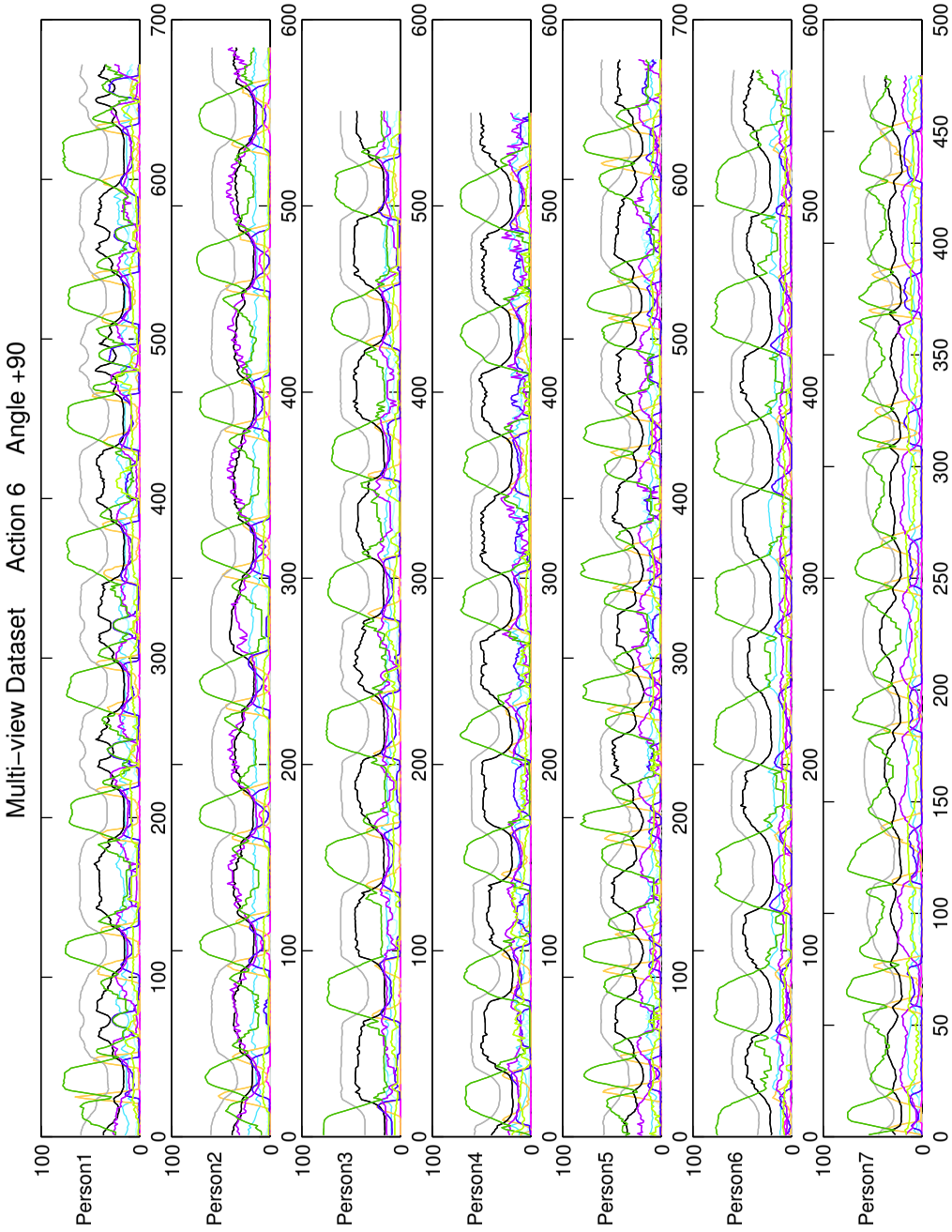

Figure A.14: Feature traces from Action 6 viewed at an angle of 90 degrees showing plots for 7 people performing a number of repetitions of the same action.races showing variation within the same action group.

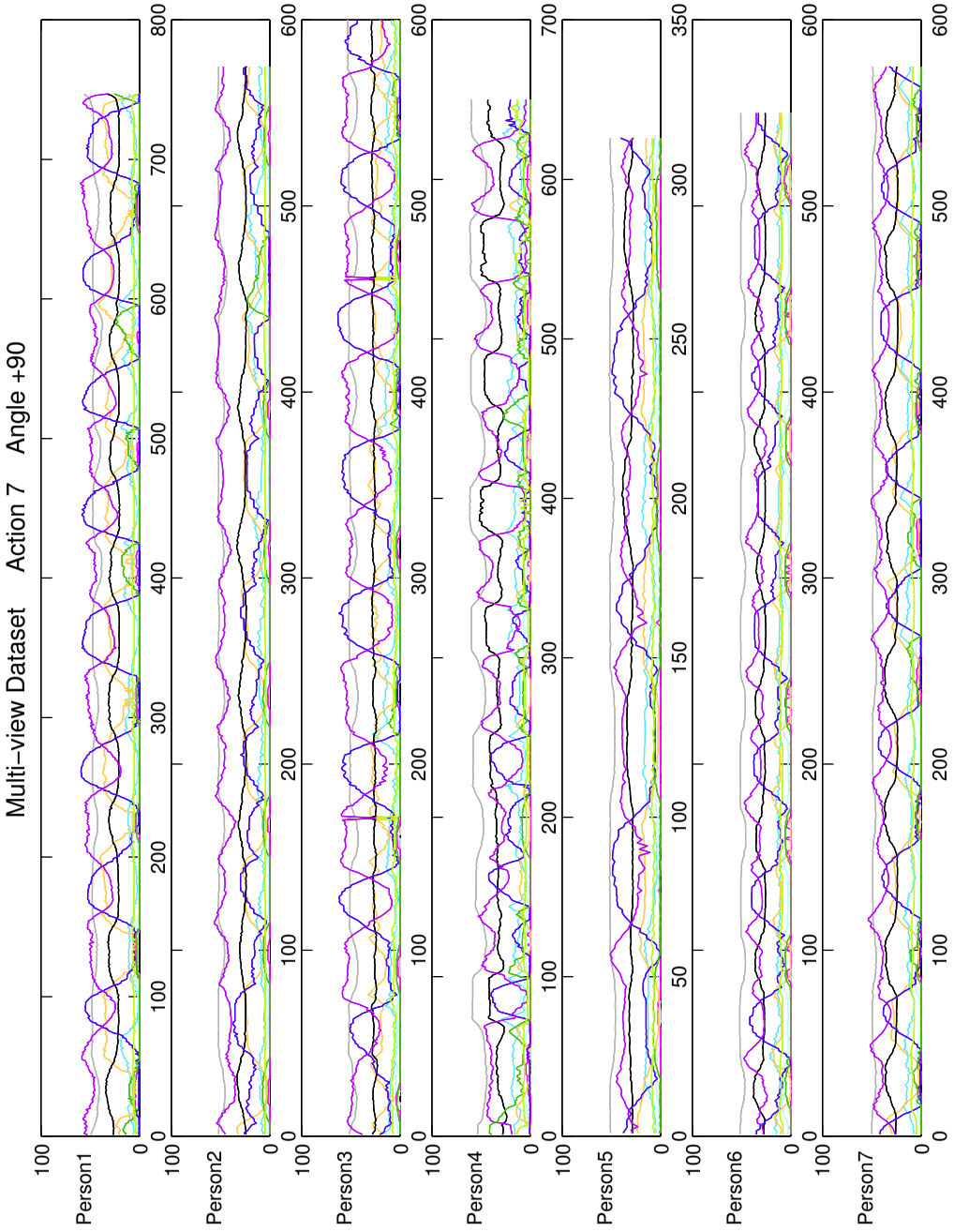

Figure A.15: Feature traces from Action 7 viewed at an angle of 90 degrees showing plots for 7 people performing a number of repetitions of the same action.

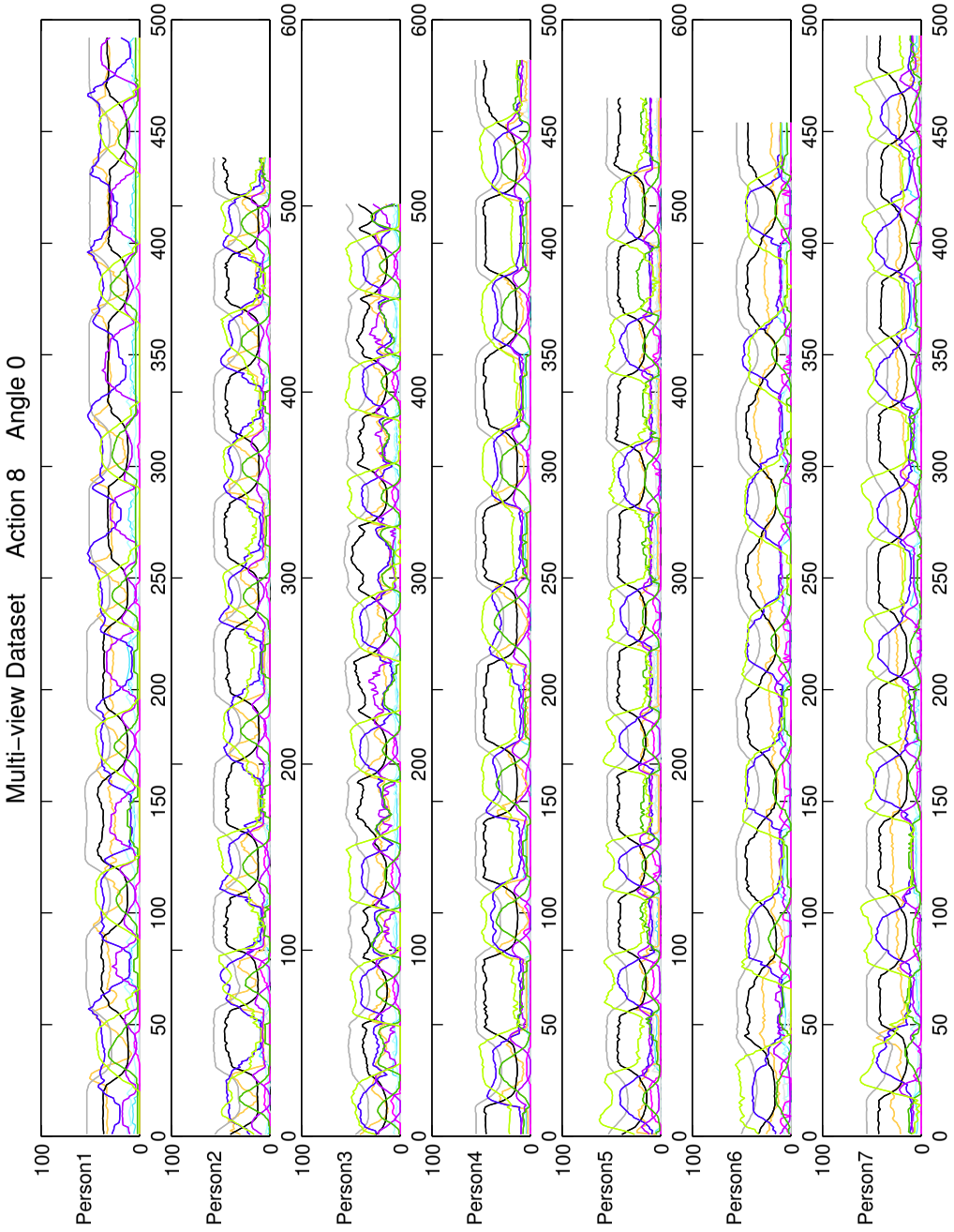

Figure A.16: Feature traces from Action 8 viewed at an angle of 0 degrees showing plots for 7 people performing a number of repetitions of the same action..

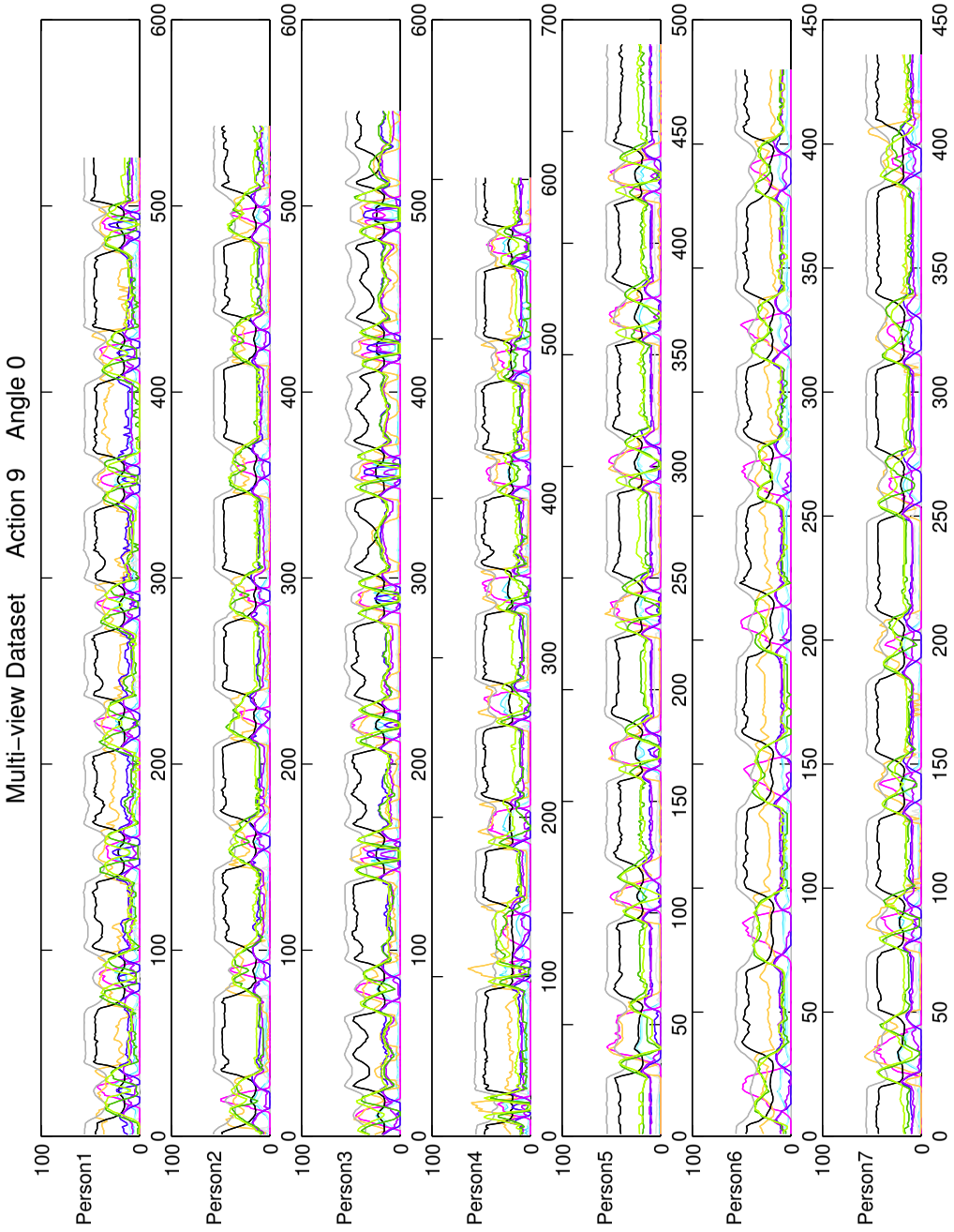

Figure A.17: Feature traces from Action 9 viewed at an angle of 0 degrees showing plots for 7 people performing a number of repetitions of the same action.

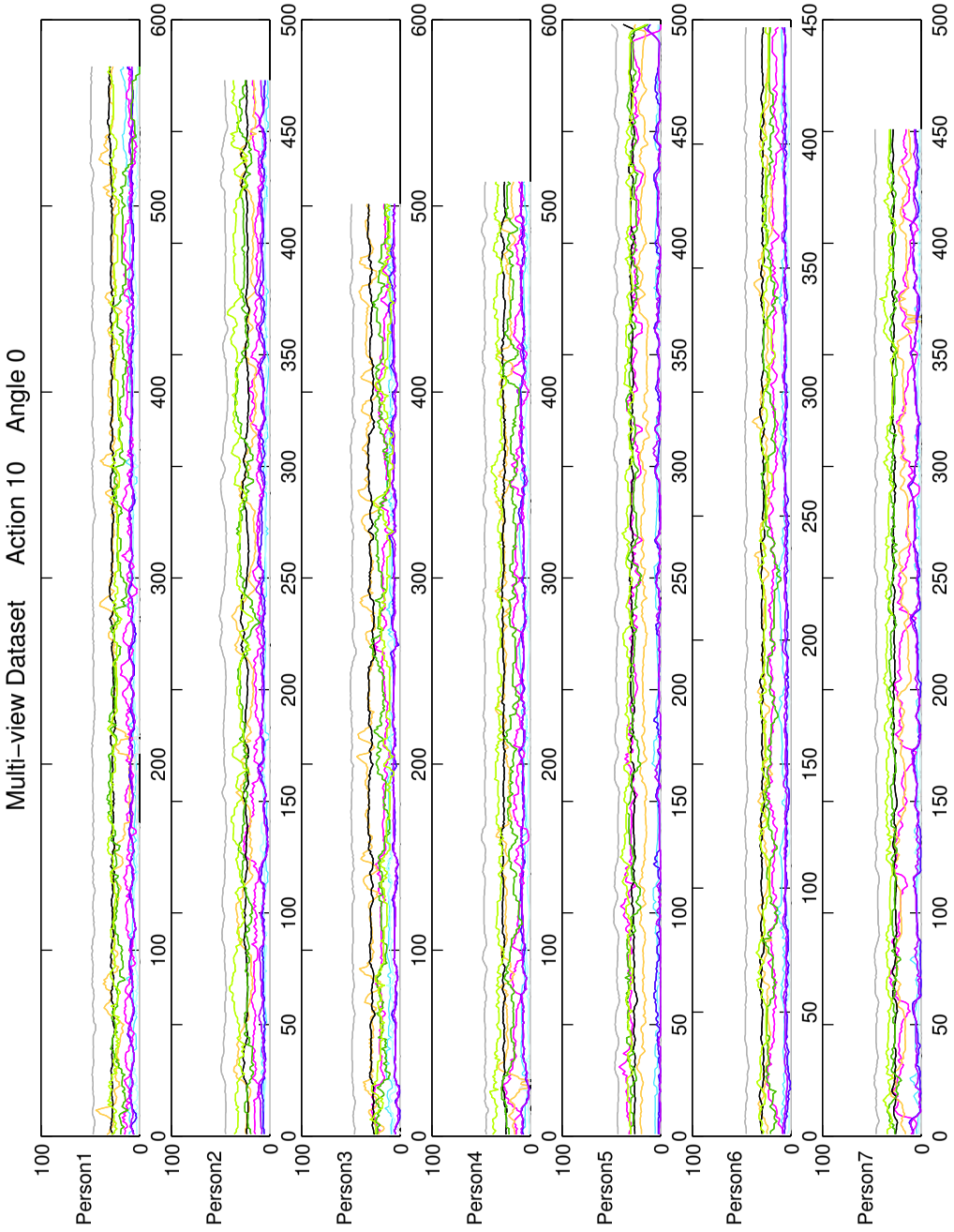

Figure A.18: Feature traces from Action 10 viewed at an angle of 0 degrees.This action was not included in the analysis as it is basically made up of one pose.

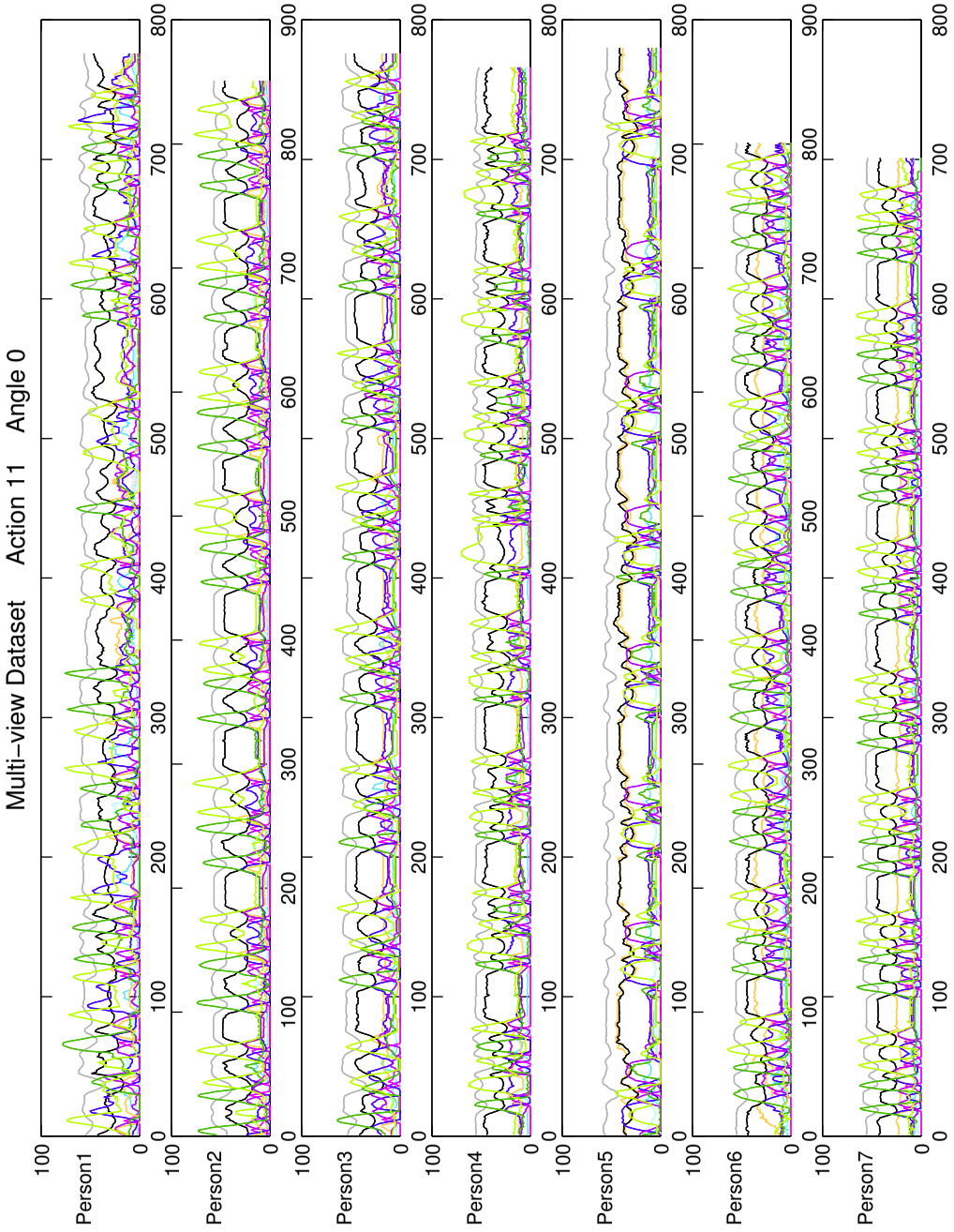

Figure A.19: Feature traces from Action 11 viewed at an angle of 0 degrees showing plots for 7 people performing a number of repetitions of the same action.

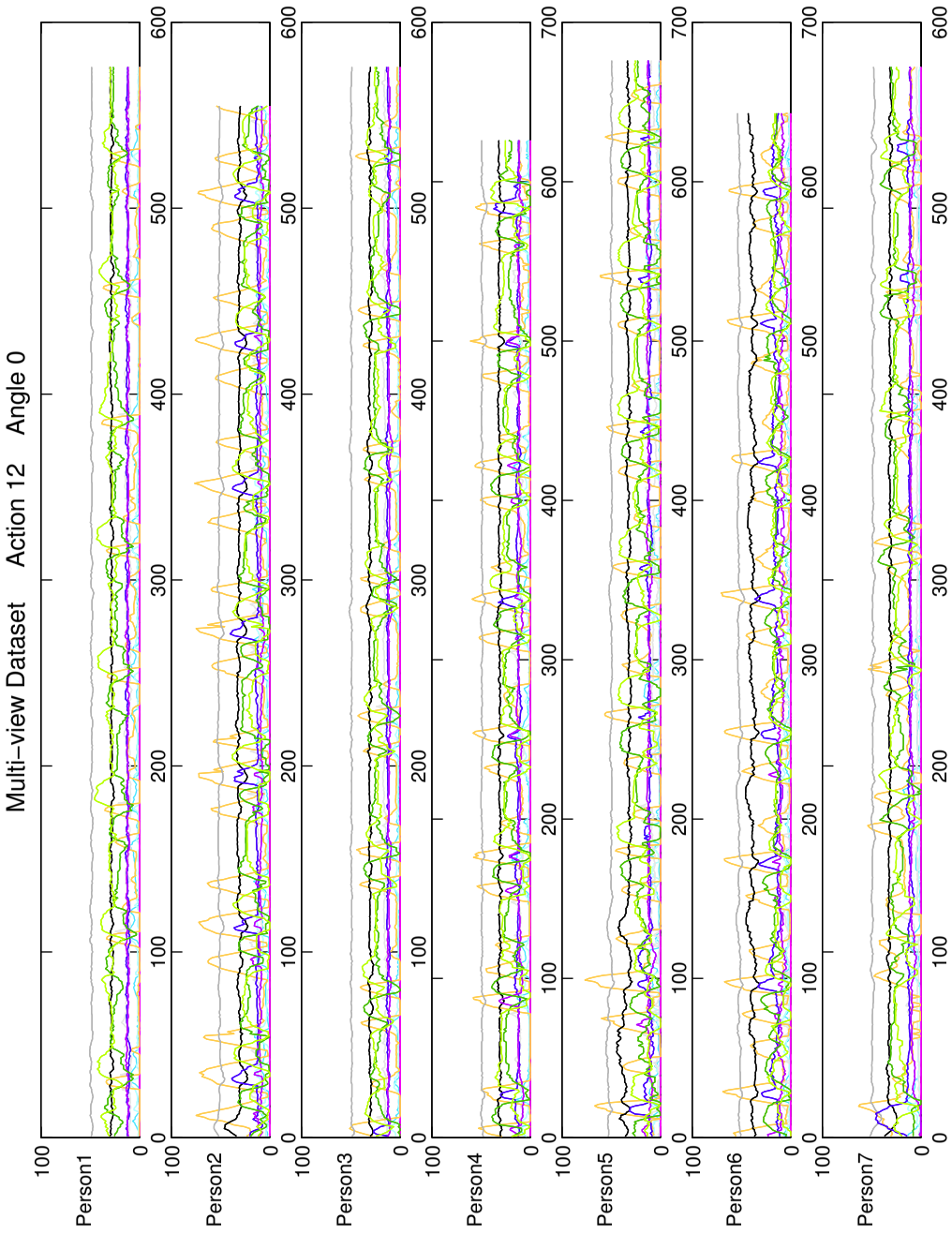

Figure A.20: Action 12 viewed at an angle of 0 degrees

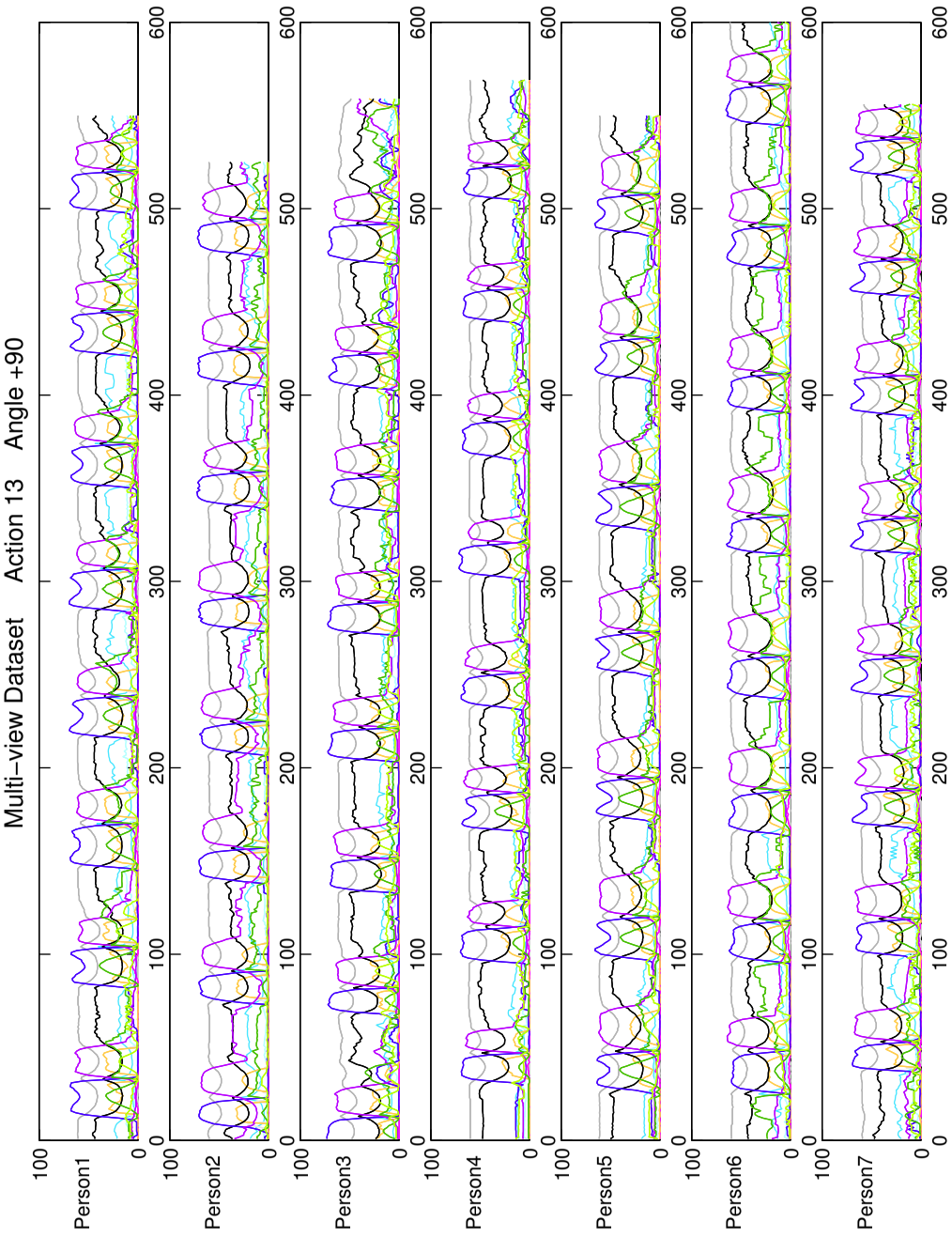

Figure A.21: Feature traces from Action 13 viewed at an angle of 90 degrees

Appendix B

Multi-view feature plots for Dataset 2

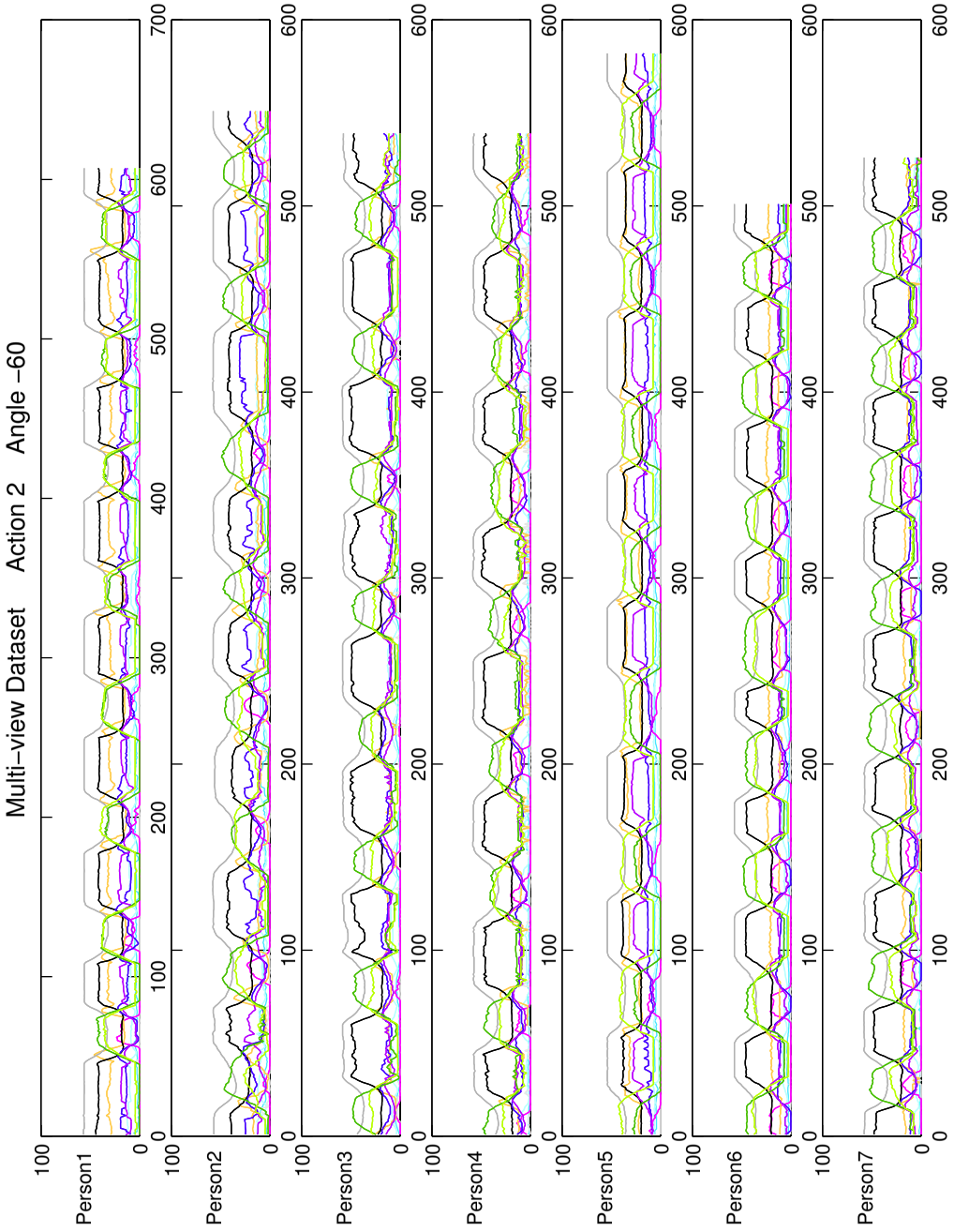

Figure B.1: Feature traces from Action 2 viewed at an angle of -60 degrees showing plots for 7 people performing a number of repetitions of the same action.

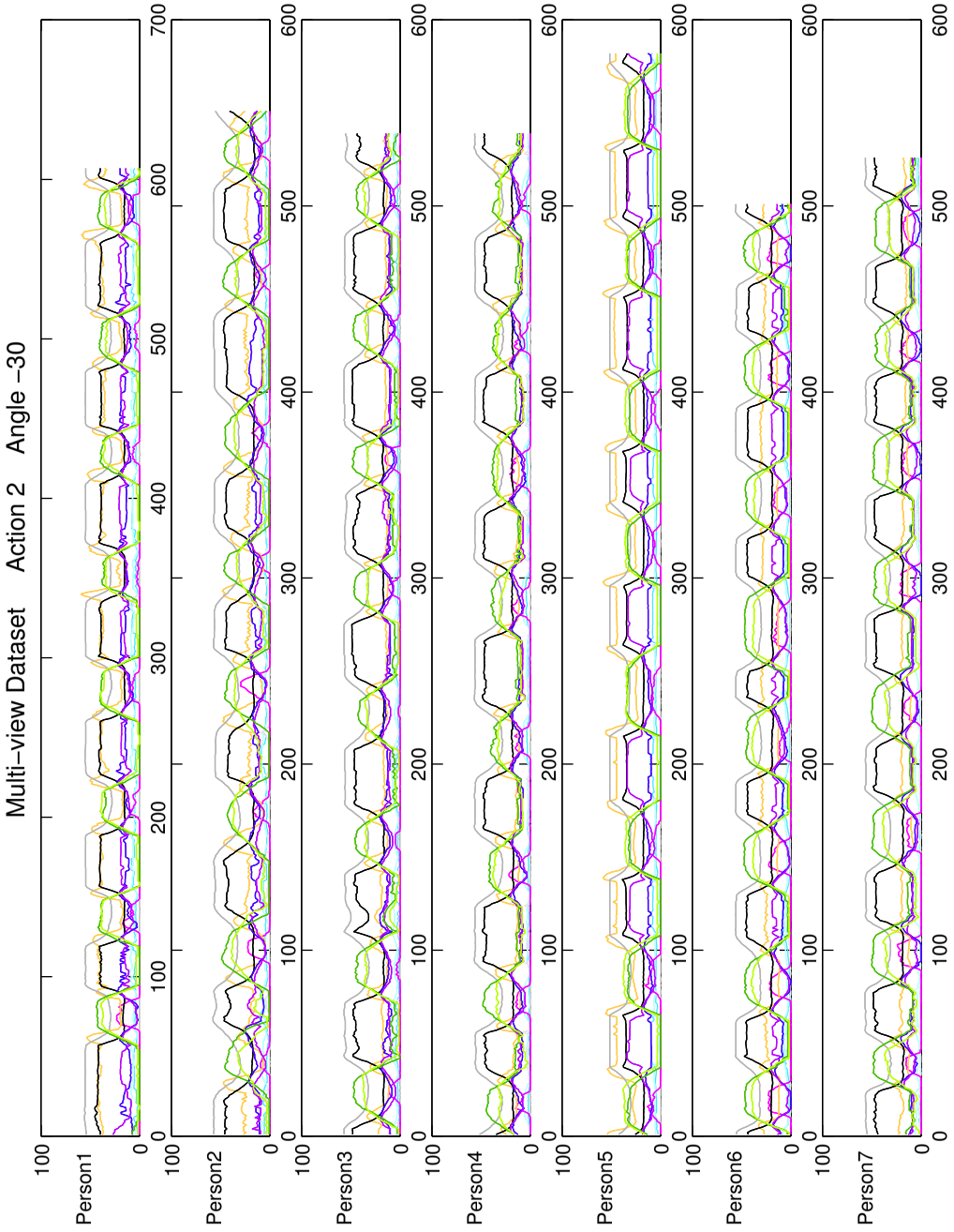

Figure B.2: Feature traces from Action 2 viewed at an angle of -30 degrees showing plots for 7 people performing a number of repetitions of the same action.

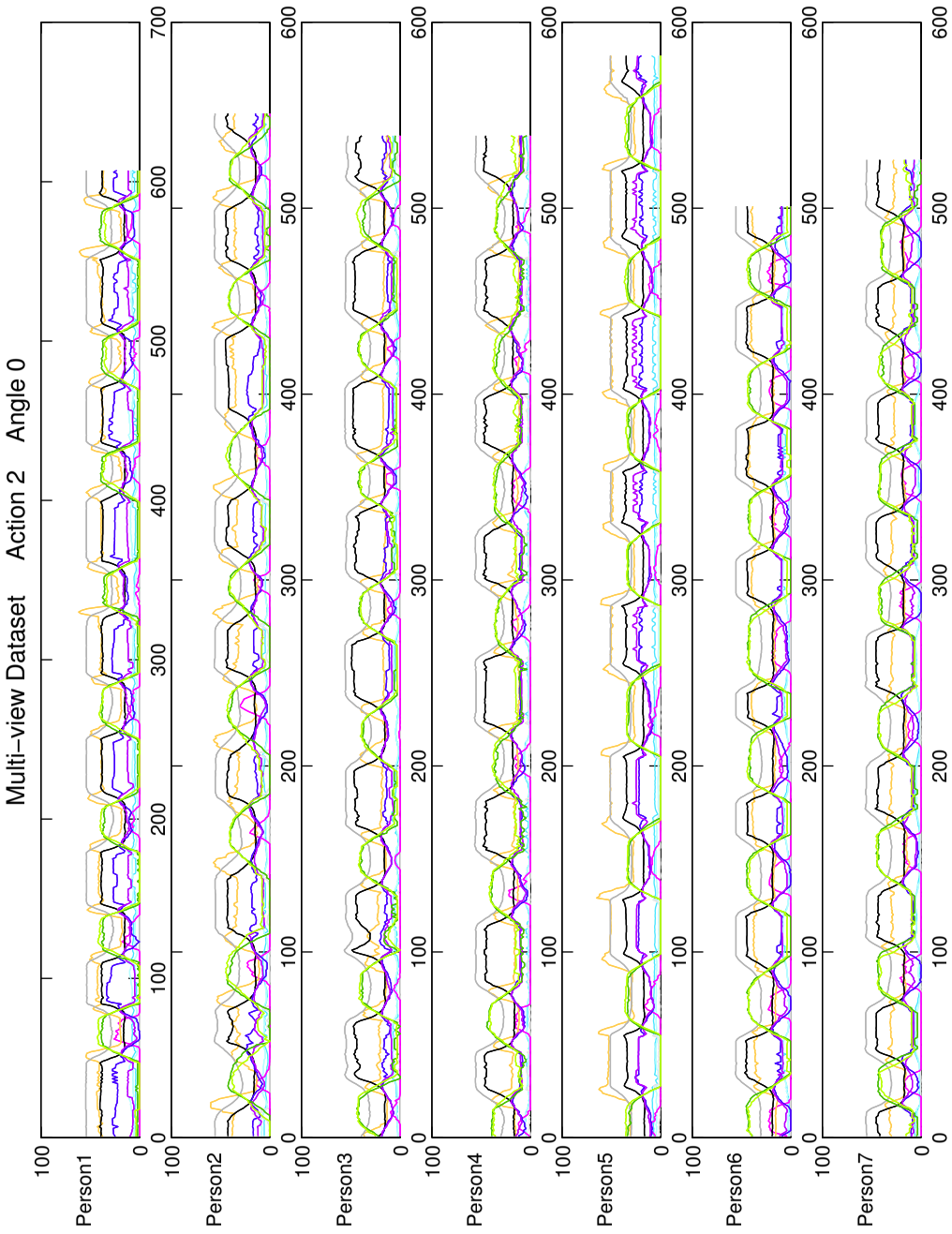

Figure B.3: Feature traces from Action 2 viewed at an angle of 0 degrees.

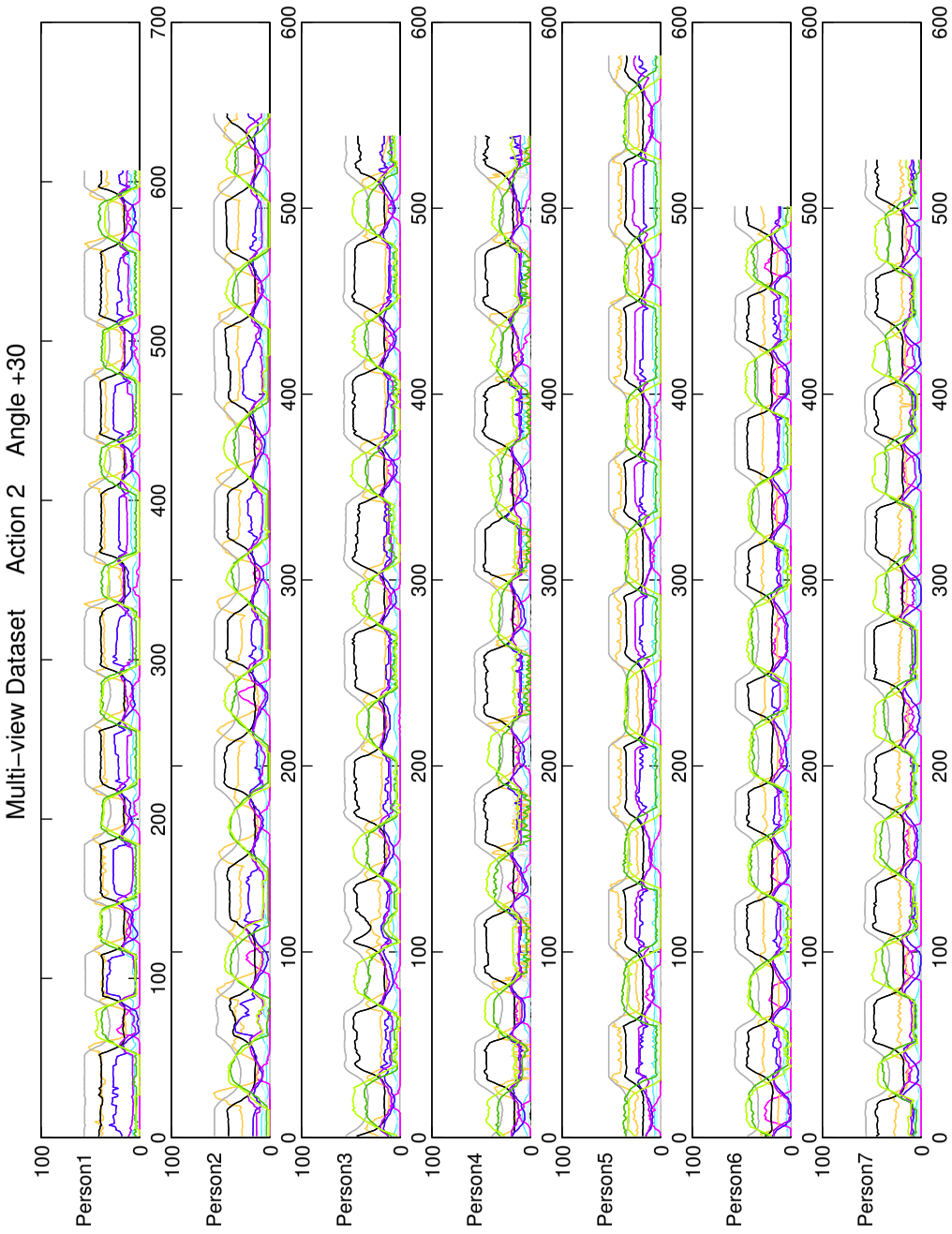

Figure B.4: Feature traces from Action 2 viewed at an angle of 30 degrees.

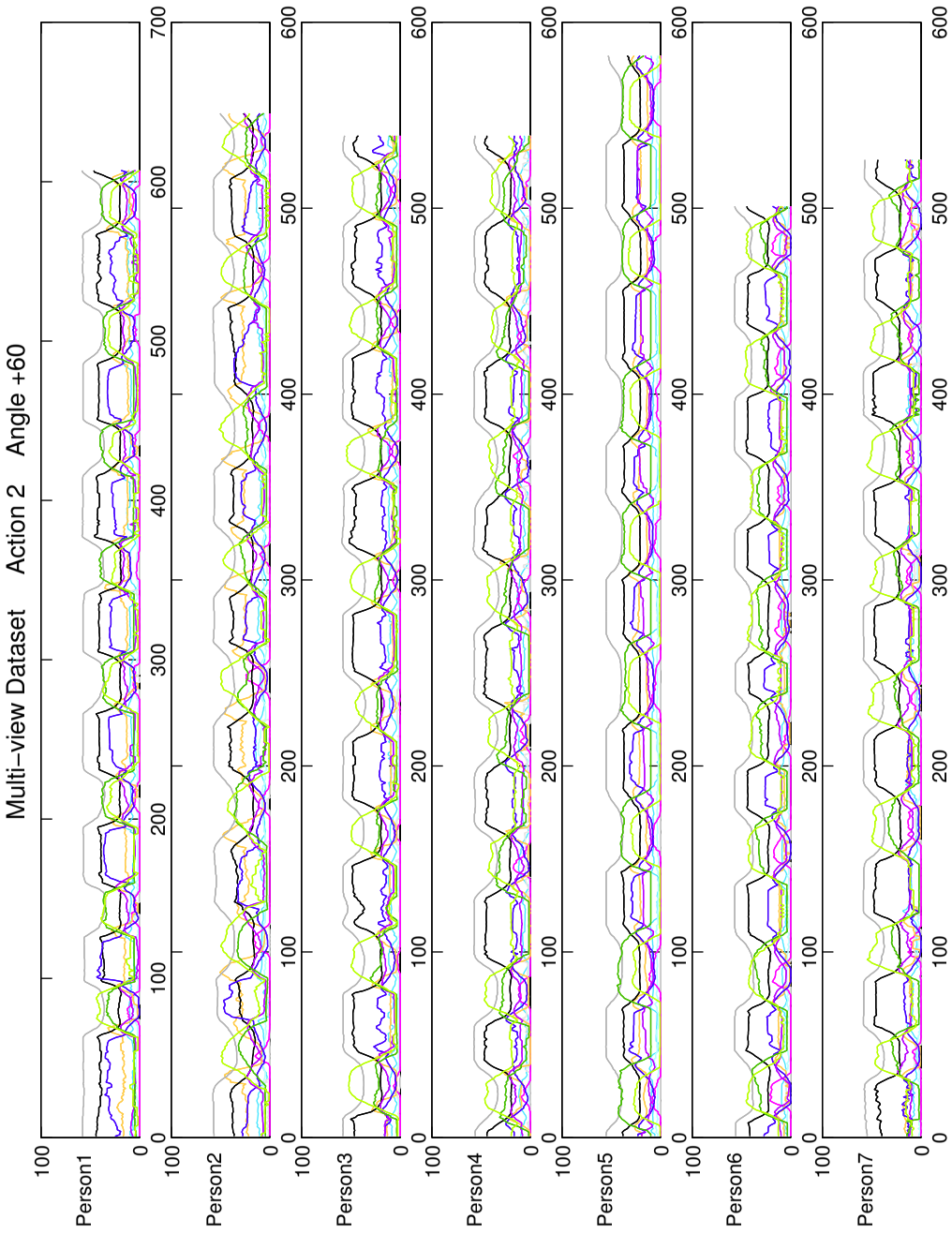

Figure B.5: Feature traces from Action 2 viewed at an angle of 60 degrees.

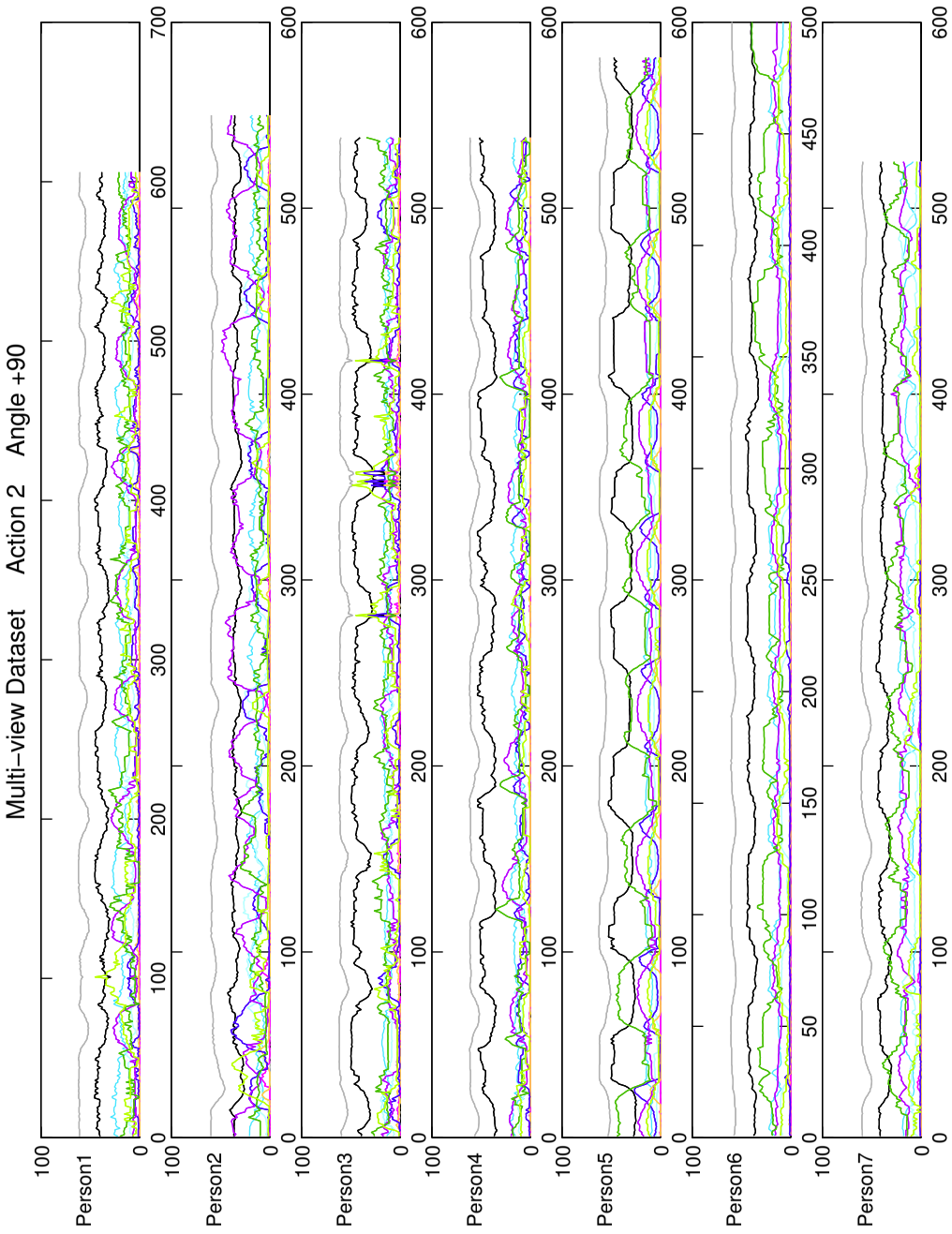

Figure B.6: Feature traces from Action 2 viewed at an angle of 90 degrees.

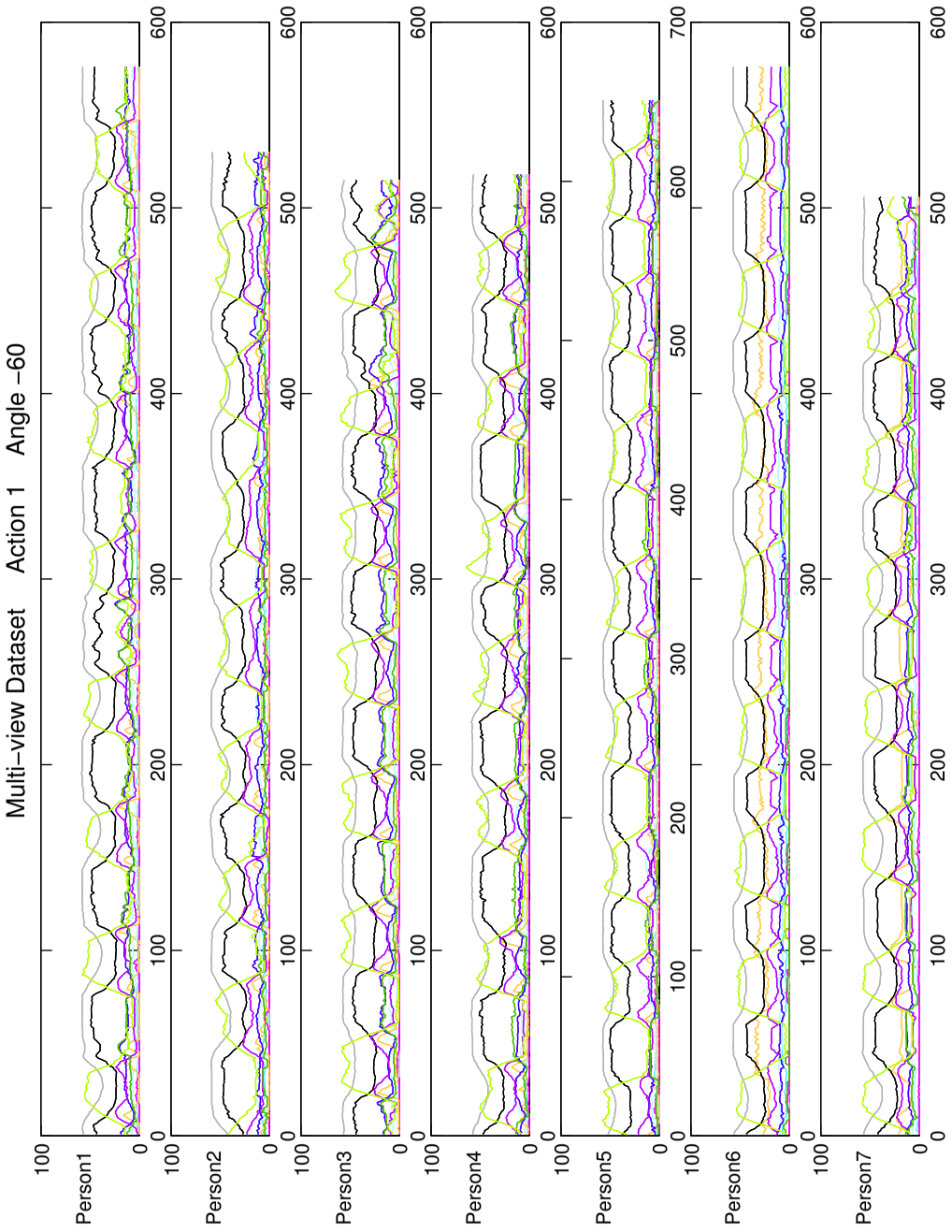

Figure B.7: Feature traces from Action 1 viewed at an angle of -60 degrees.

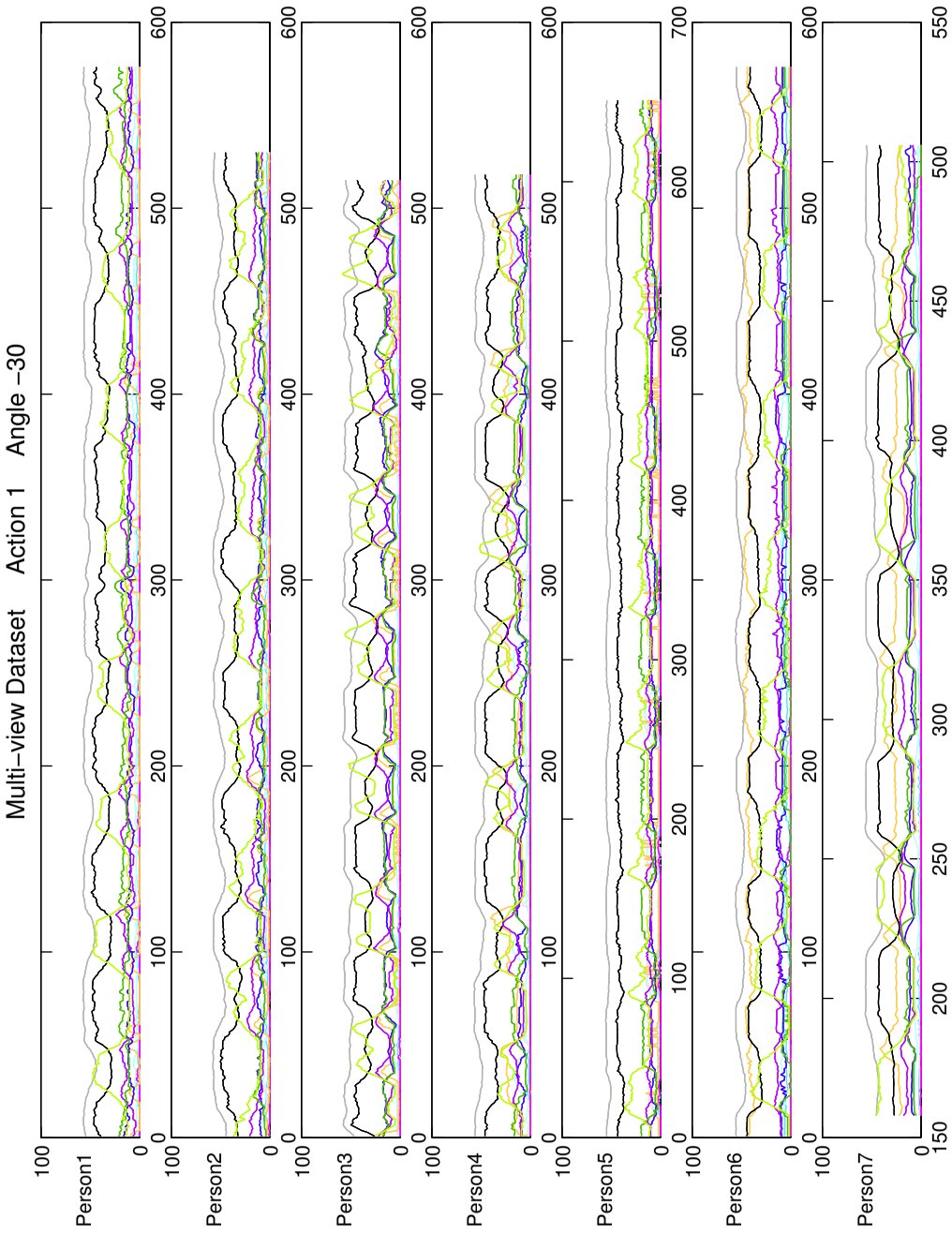

Figure B.8: Feature traces from Action 1 viewed at an angle of -30 degrees..

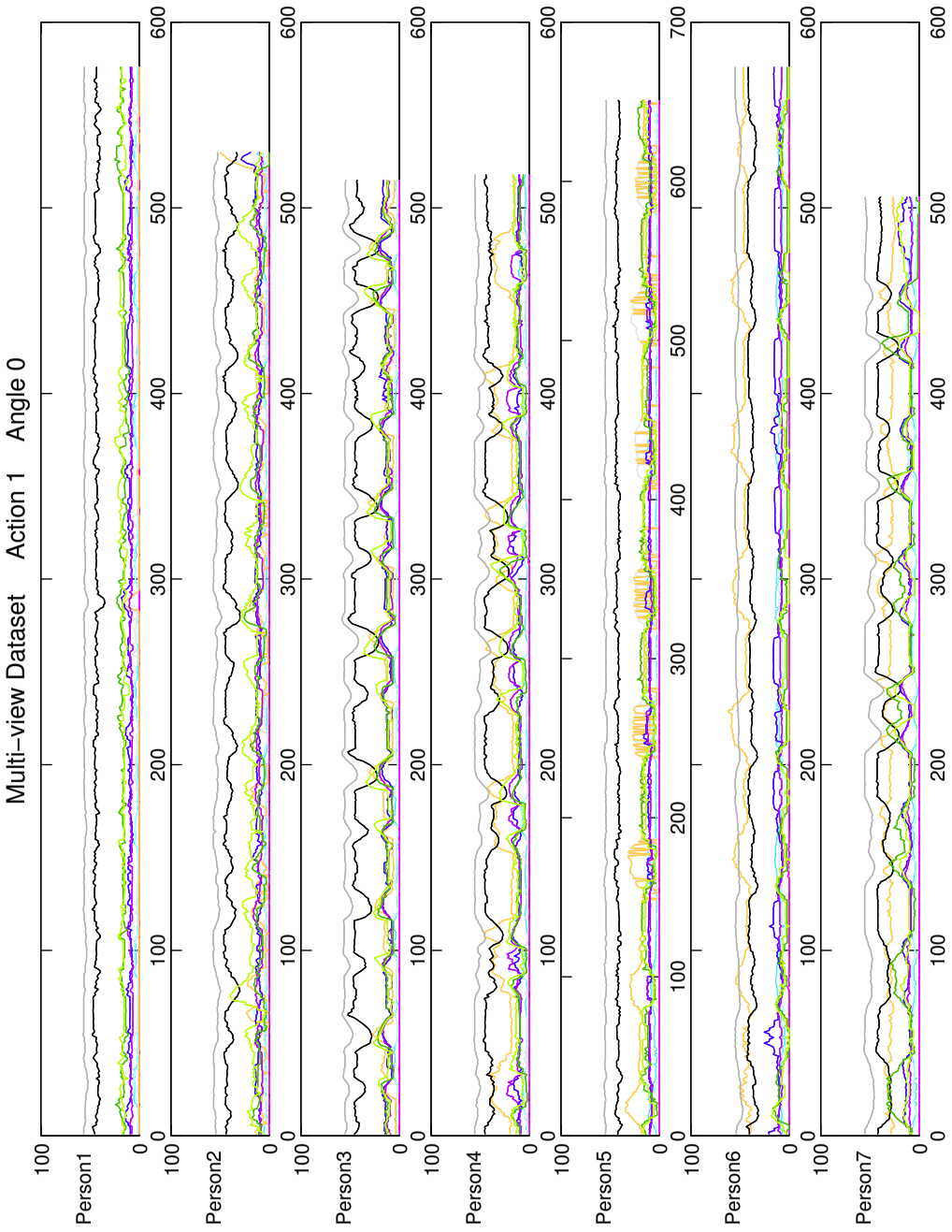

Figure B.9: Feature traces from Action 1 viewed at an angle of 0 degrees.

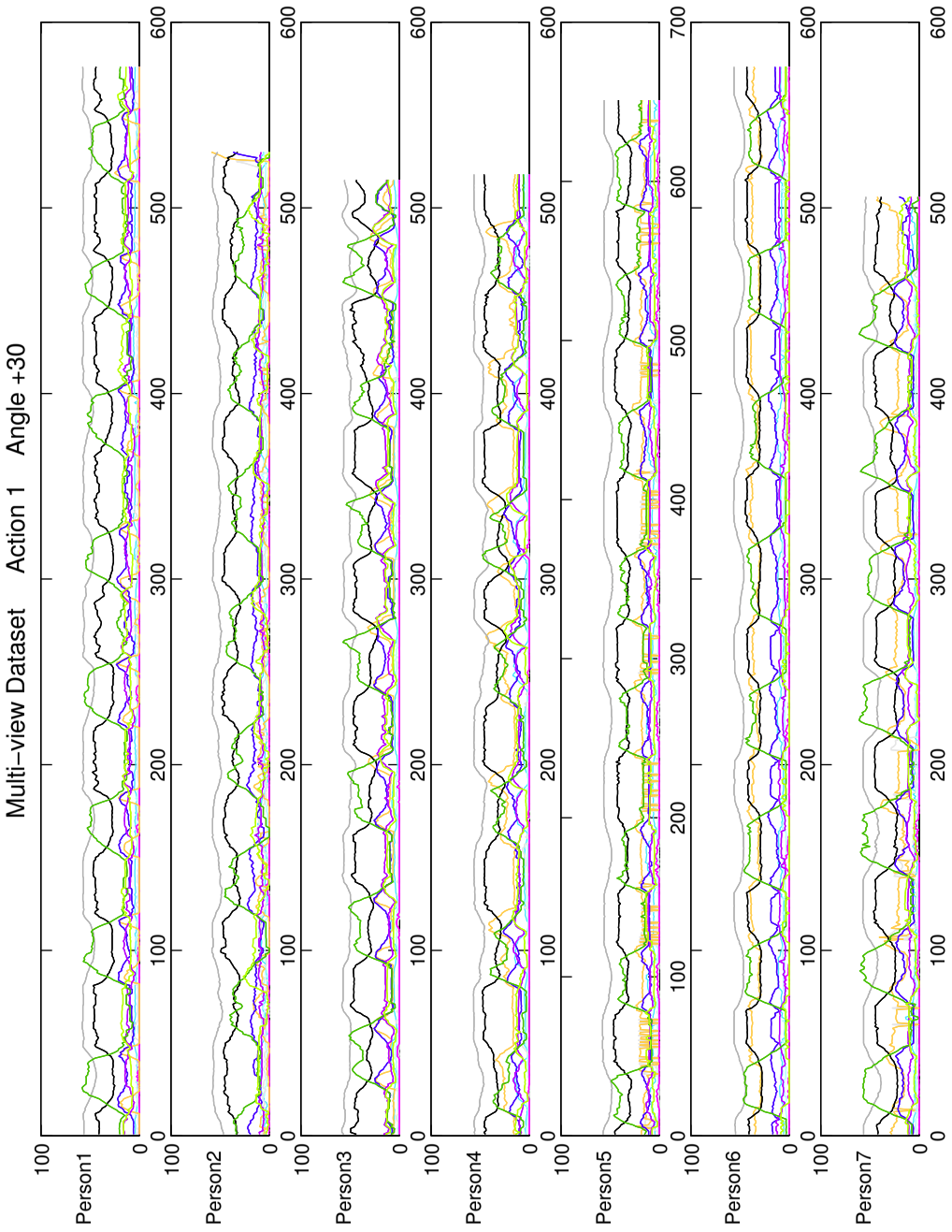

Figure B.10: Feature traces from Action 1 viewed at an angle of 30 degrees.

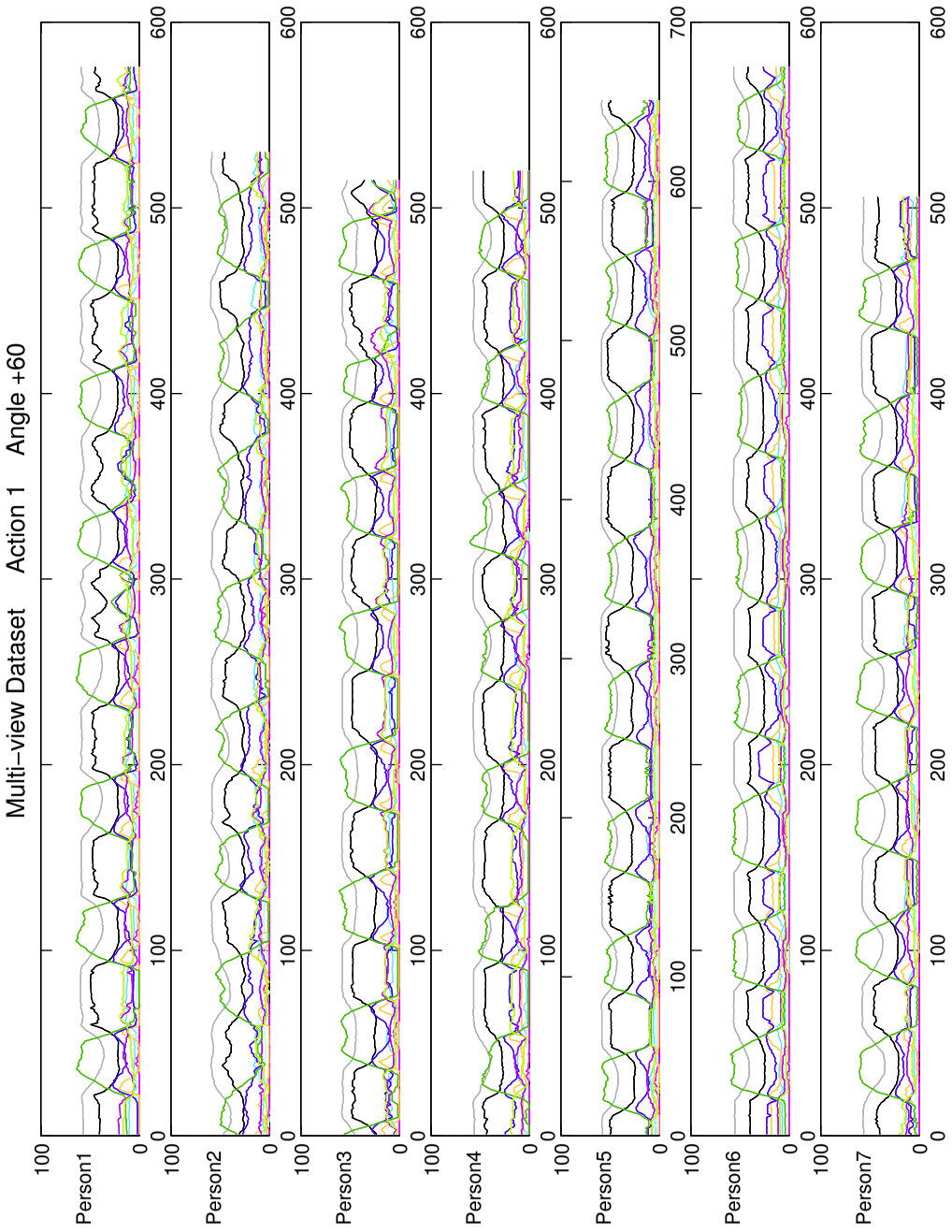

Figure B.11: Feature traces from Action 1 viewed at an angle of 60 degrees.

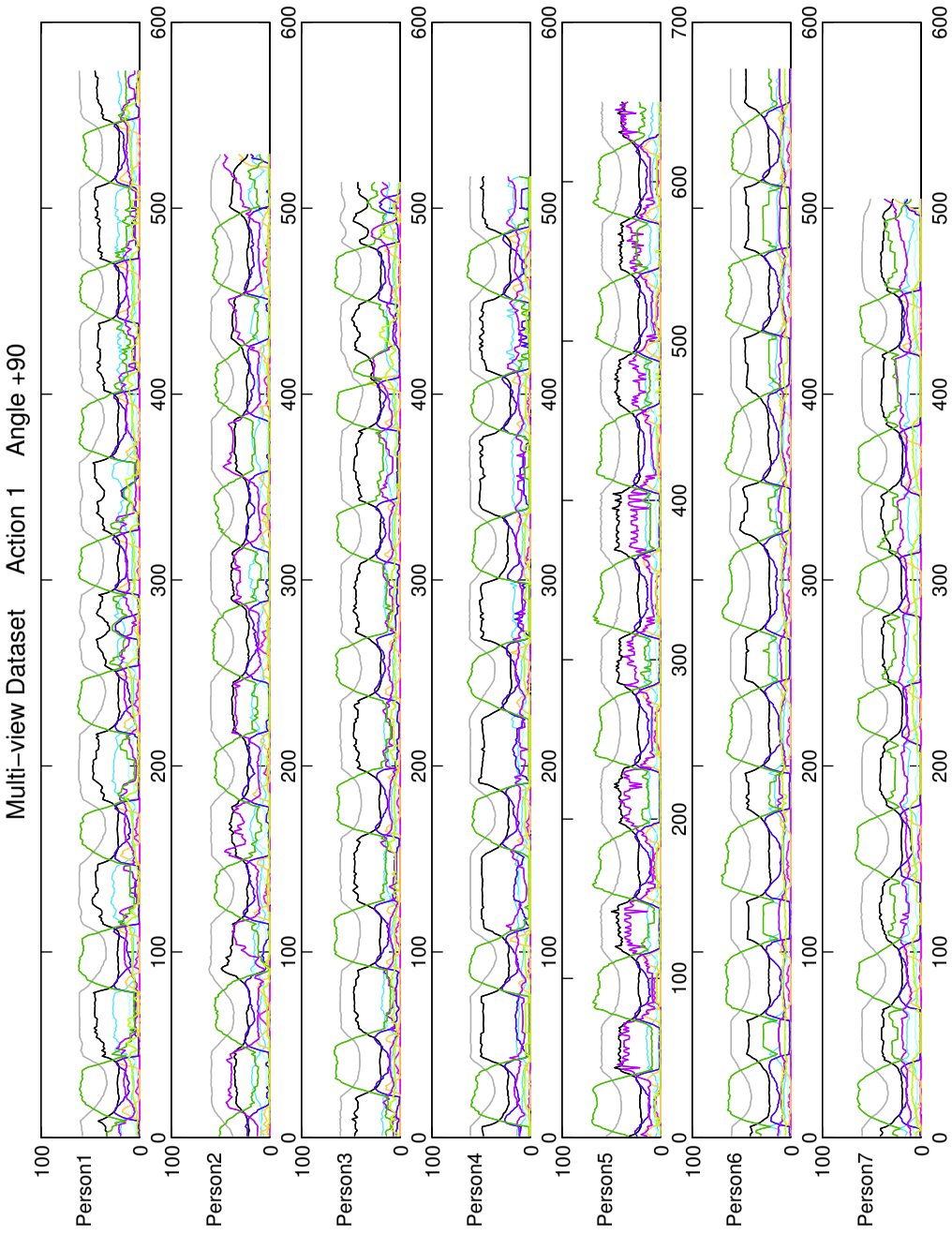

Figure B.12: Feature traces from Action 1 viewed at an angle of 90 degrees.

Appendix C

Multi-view feature plots after clustering

The plots show feature traces before and after clustering for the angle from which the action is best seen.

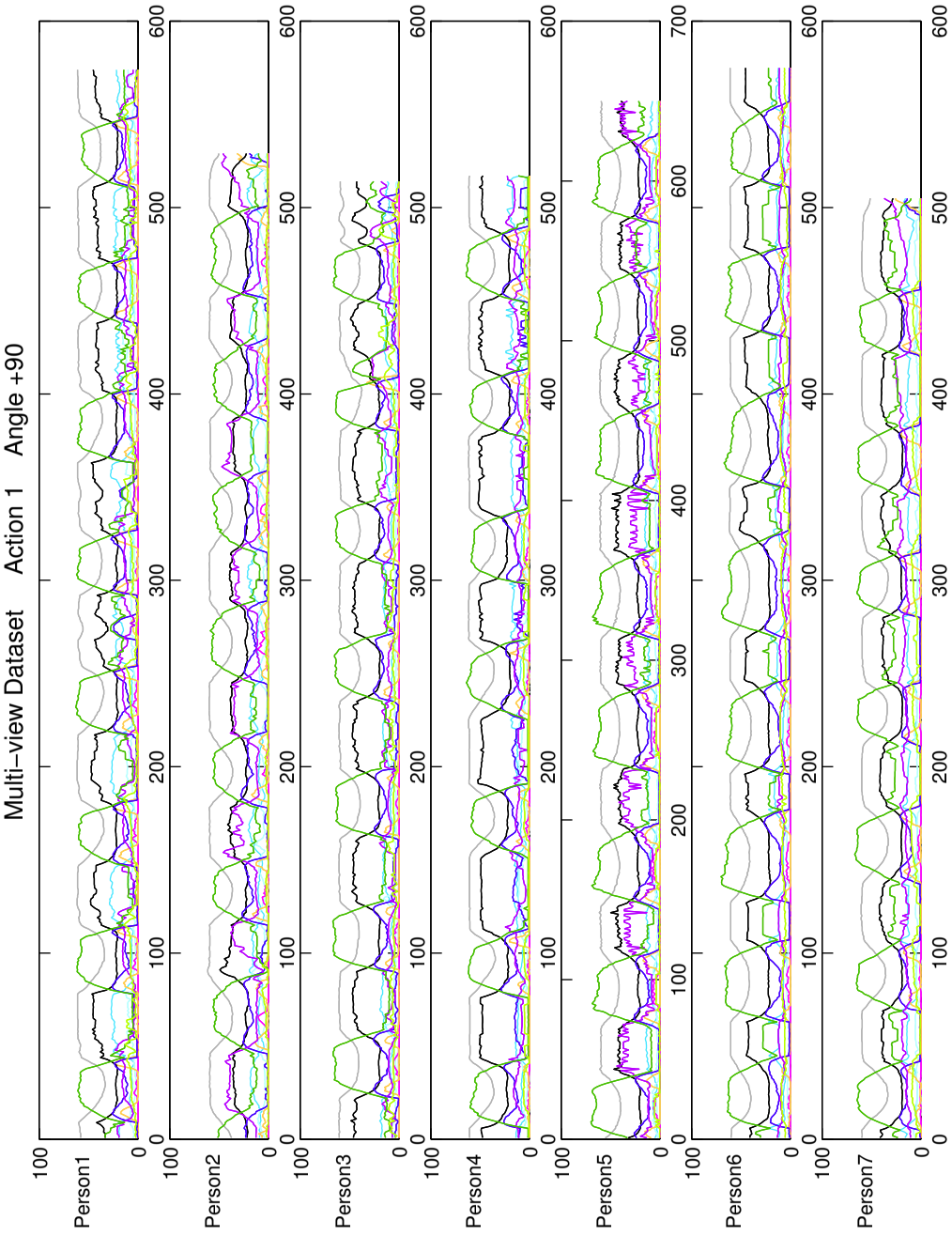

Figure C.1: The feature traces of Action 1 before processing.

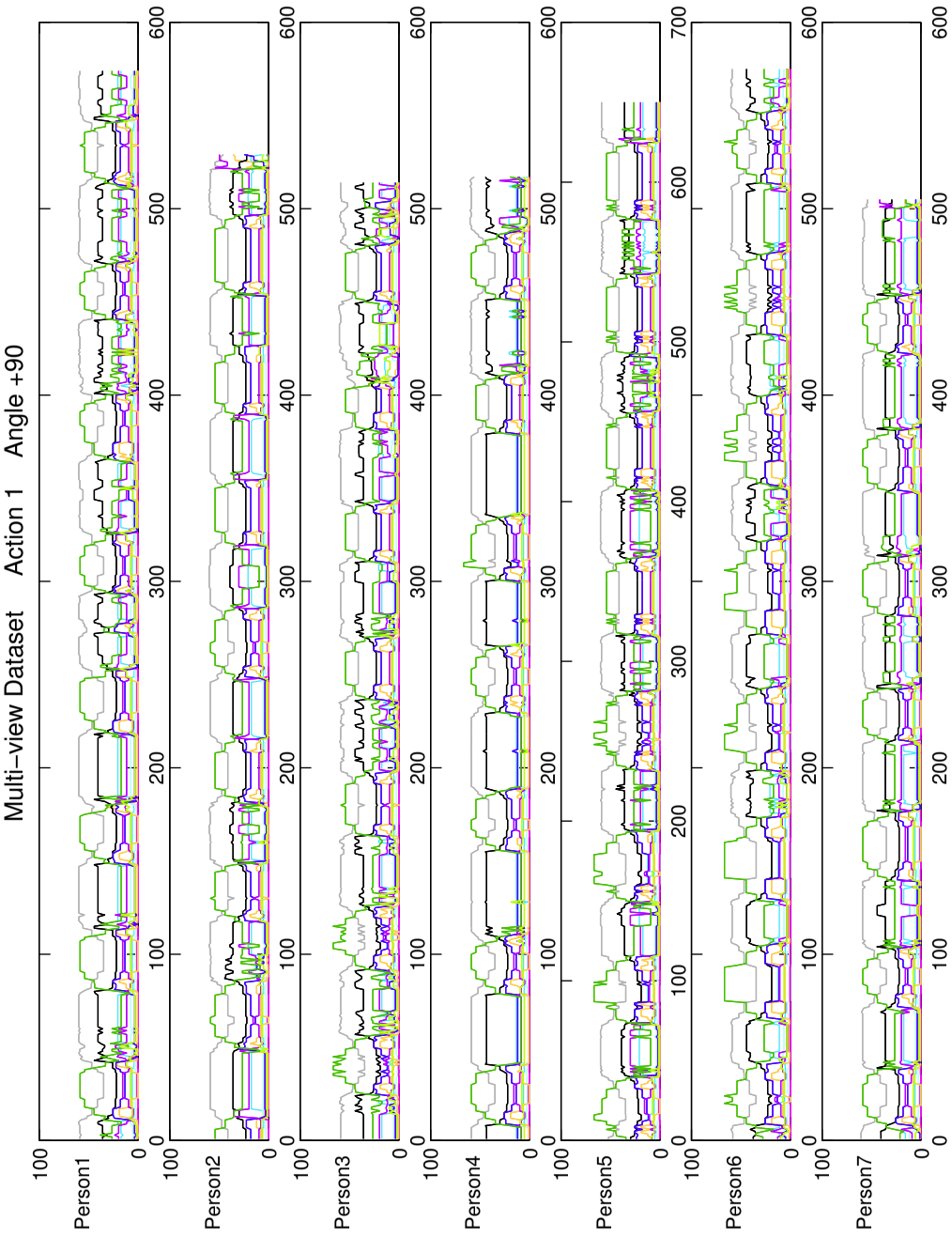

Figure C.2: The Action 1 traces after clustering

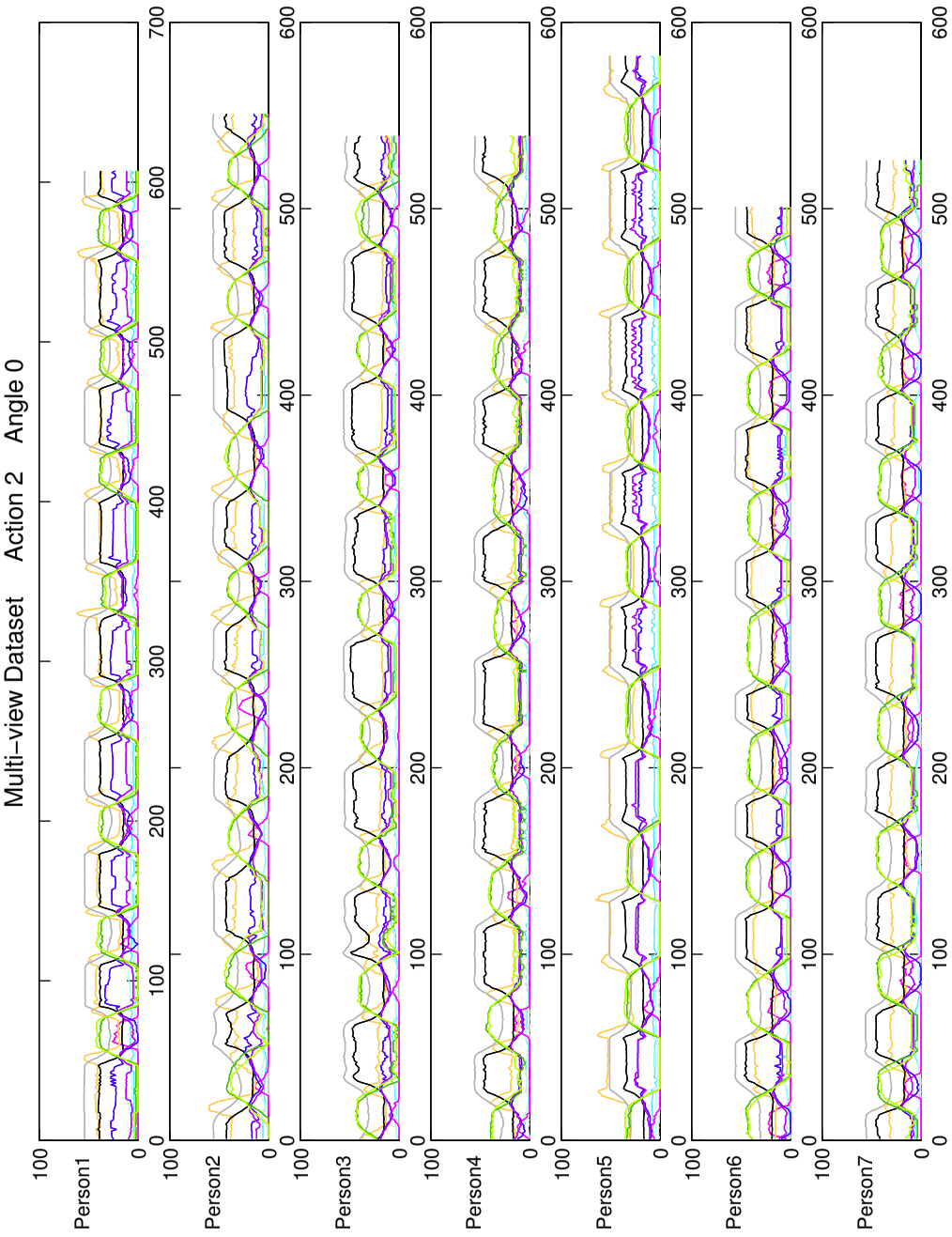

Figure C.3: The feature traces of Action 2 before processing.

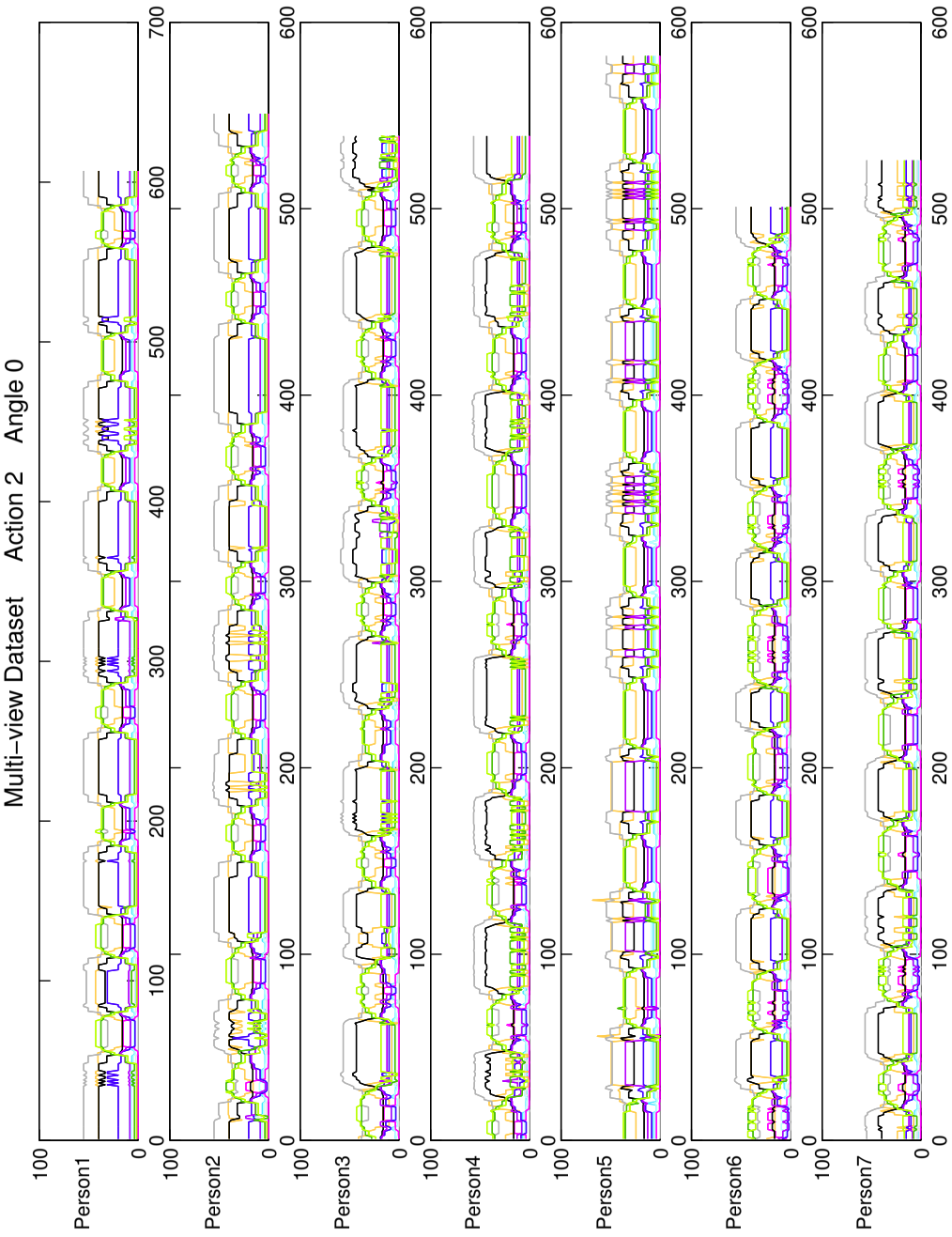

Figure C.4: The Action 2 traces after clustering

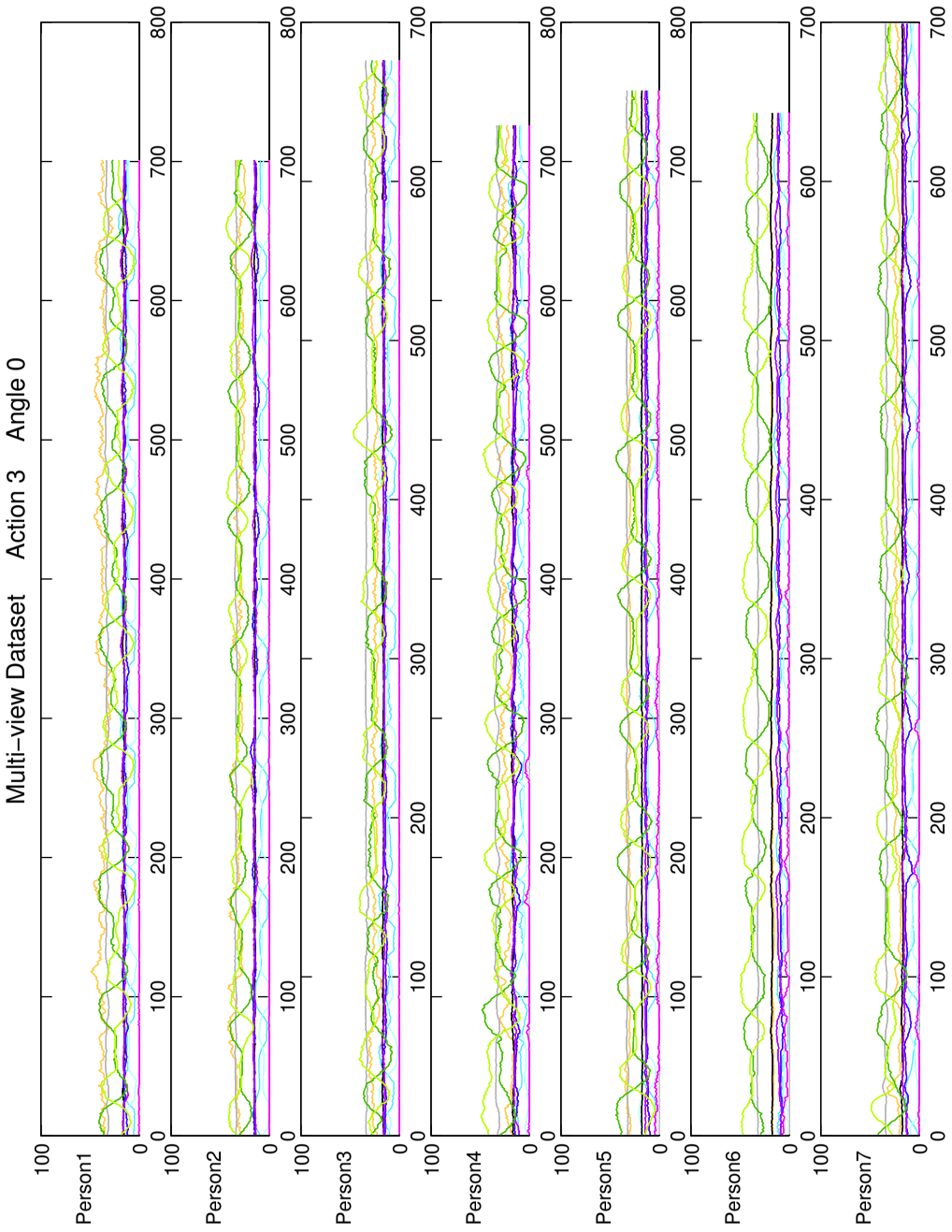

Figure C.5: The feature traces of Action 3 before processing.

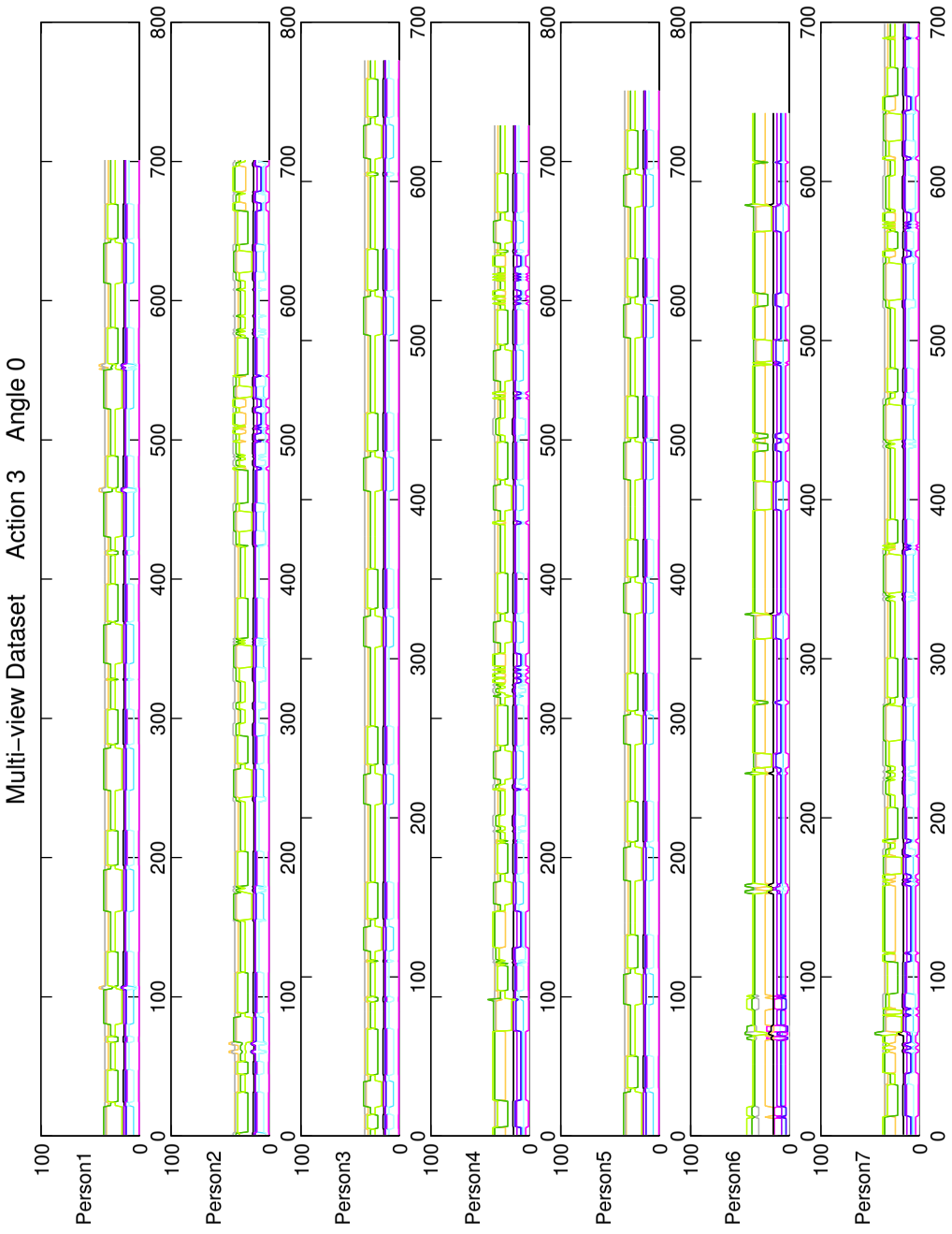

Figure C.6: The Action 3 traces after clustering

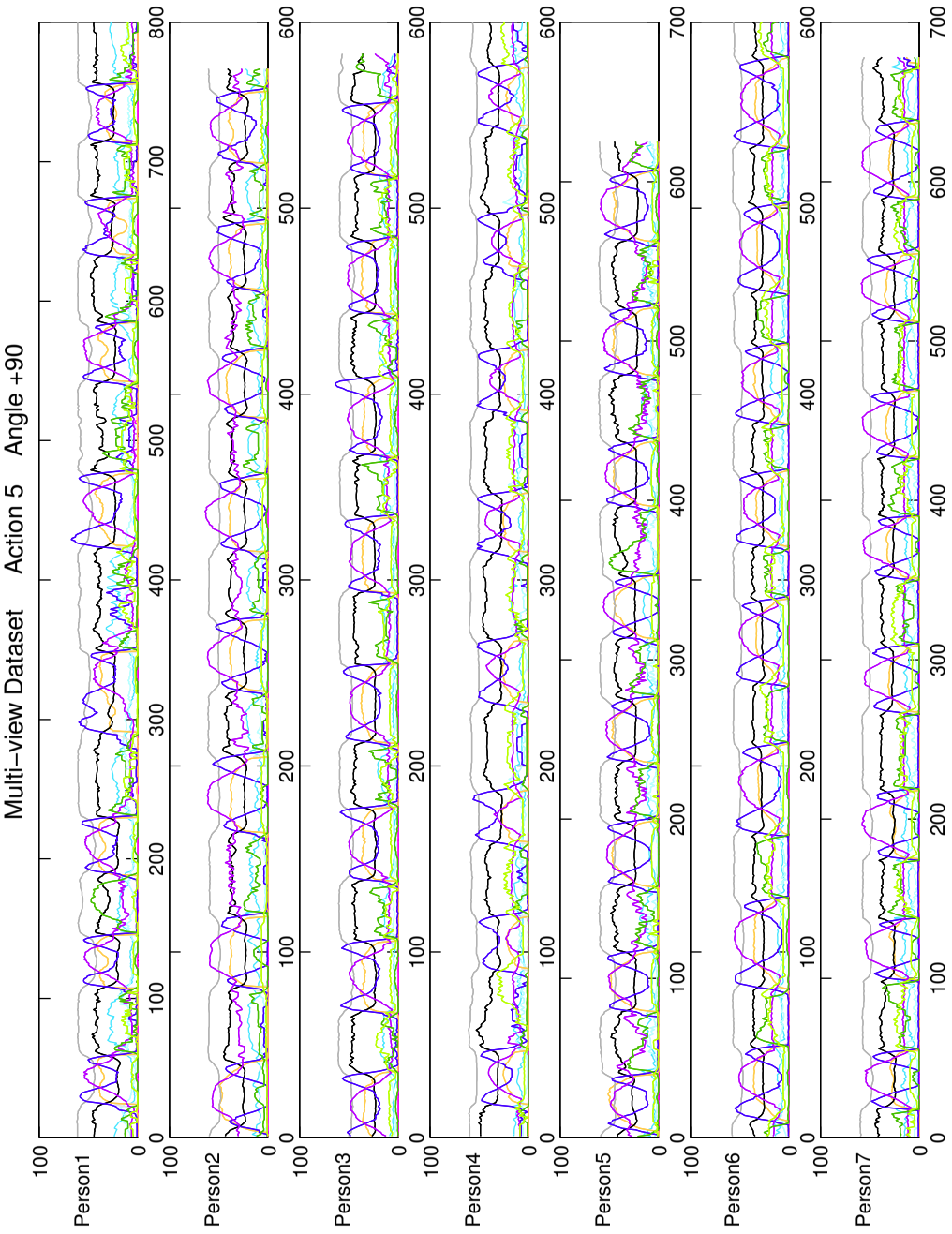

Figure C.7: The feature traces of Action 5 before processing.

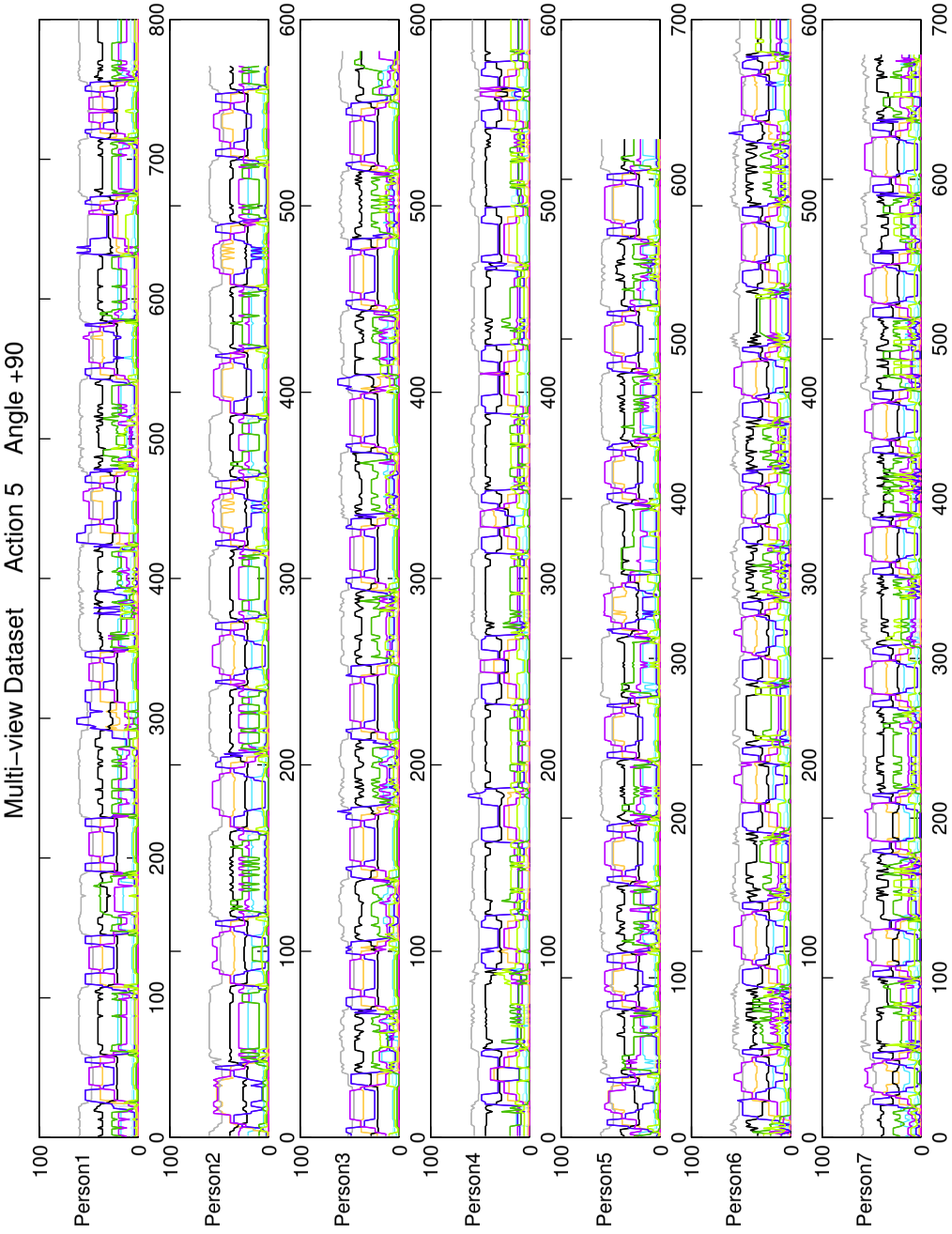

Figure C.8: The Action 5 traces after clustering

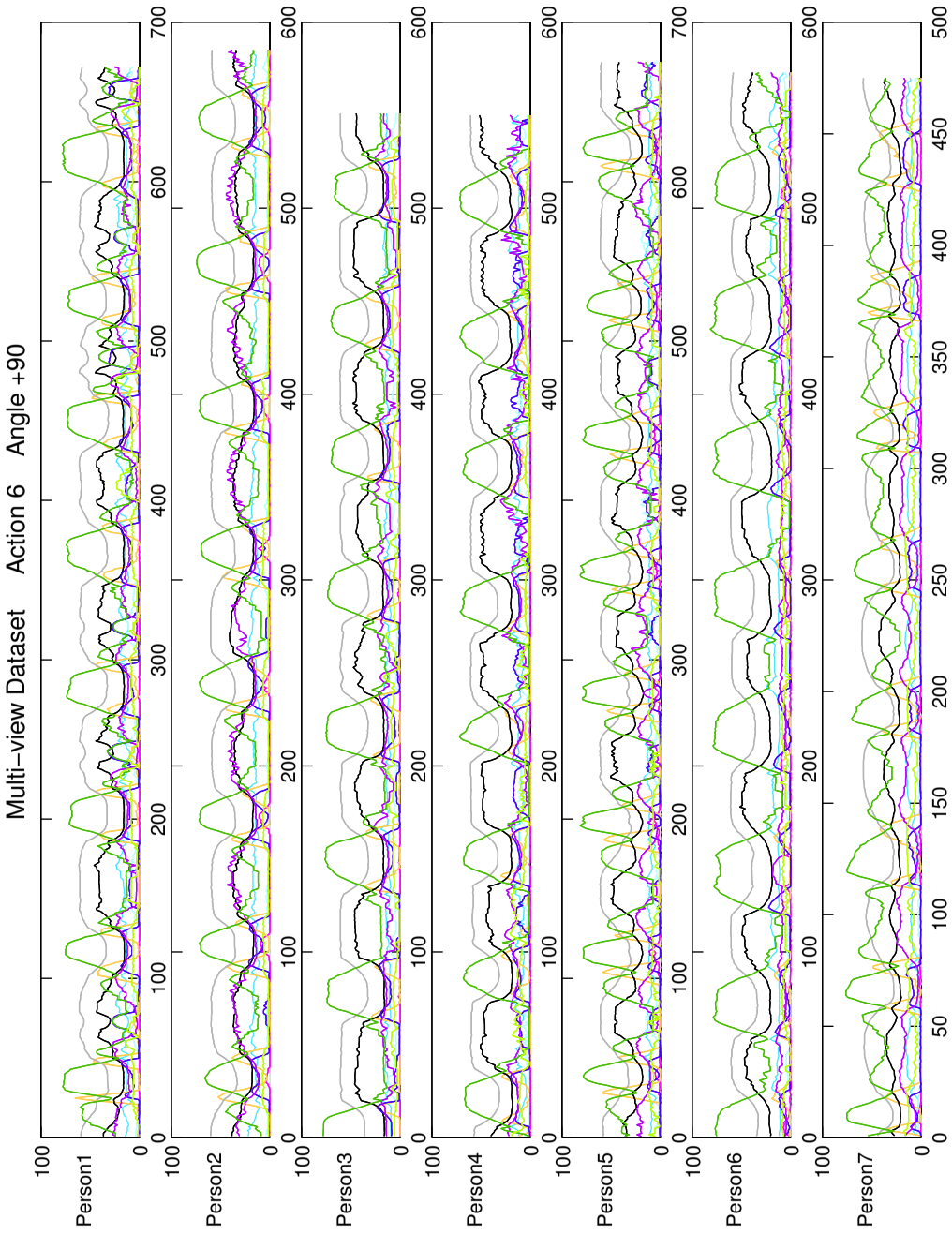

Figure C.9: The feature traces of Action 6 before processing.

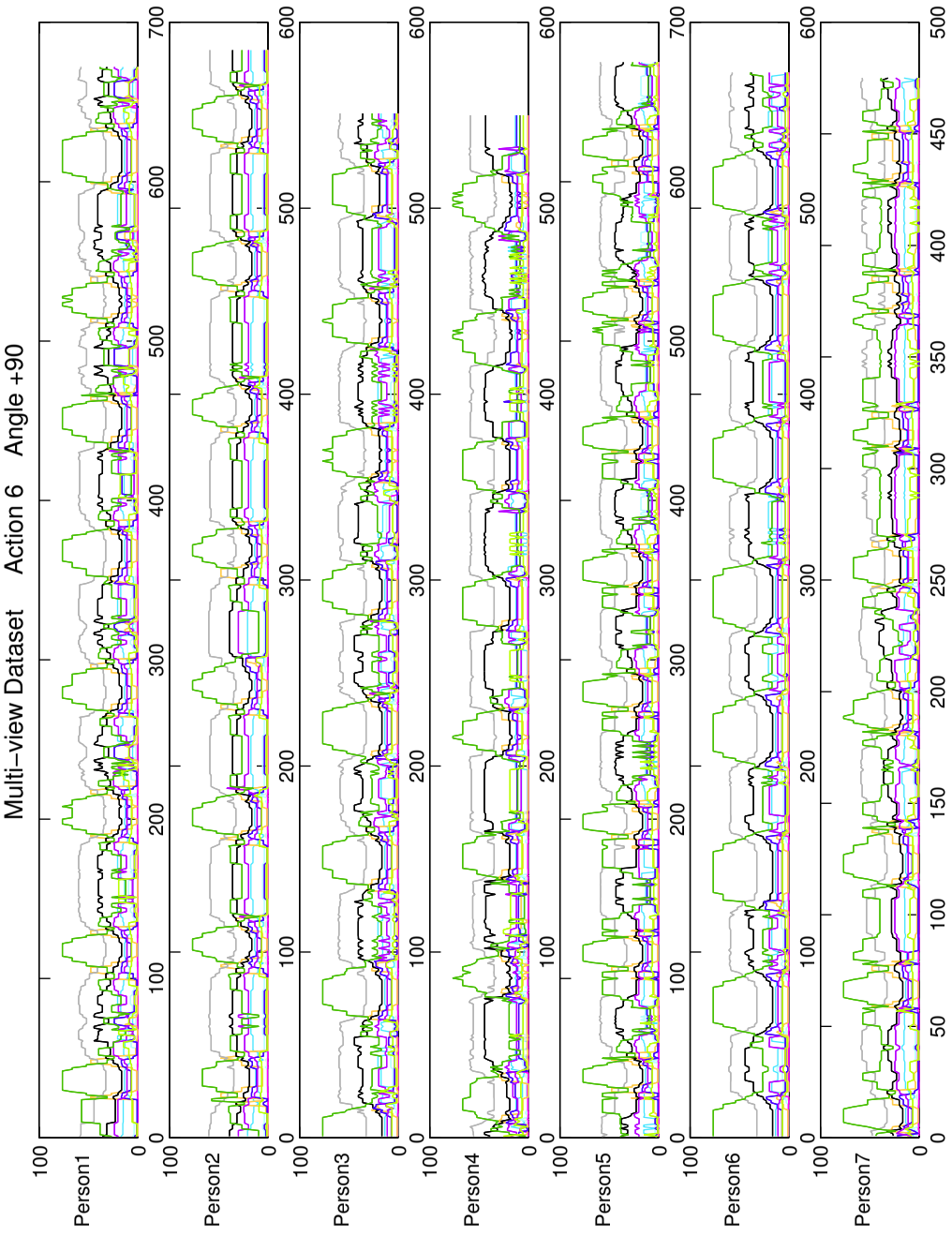

Figure C.10: The Action 6 traces after clustering

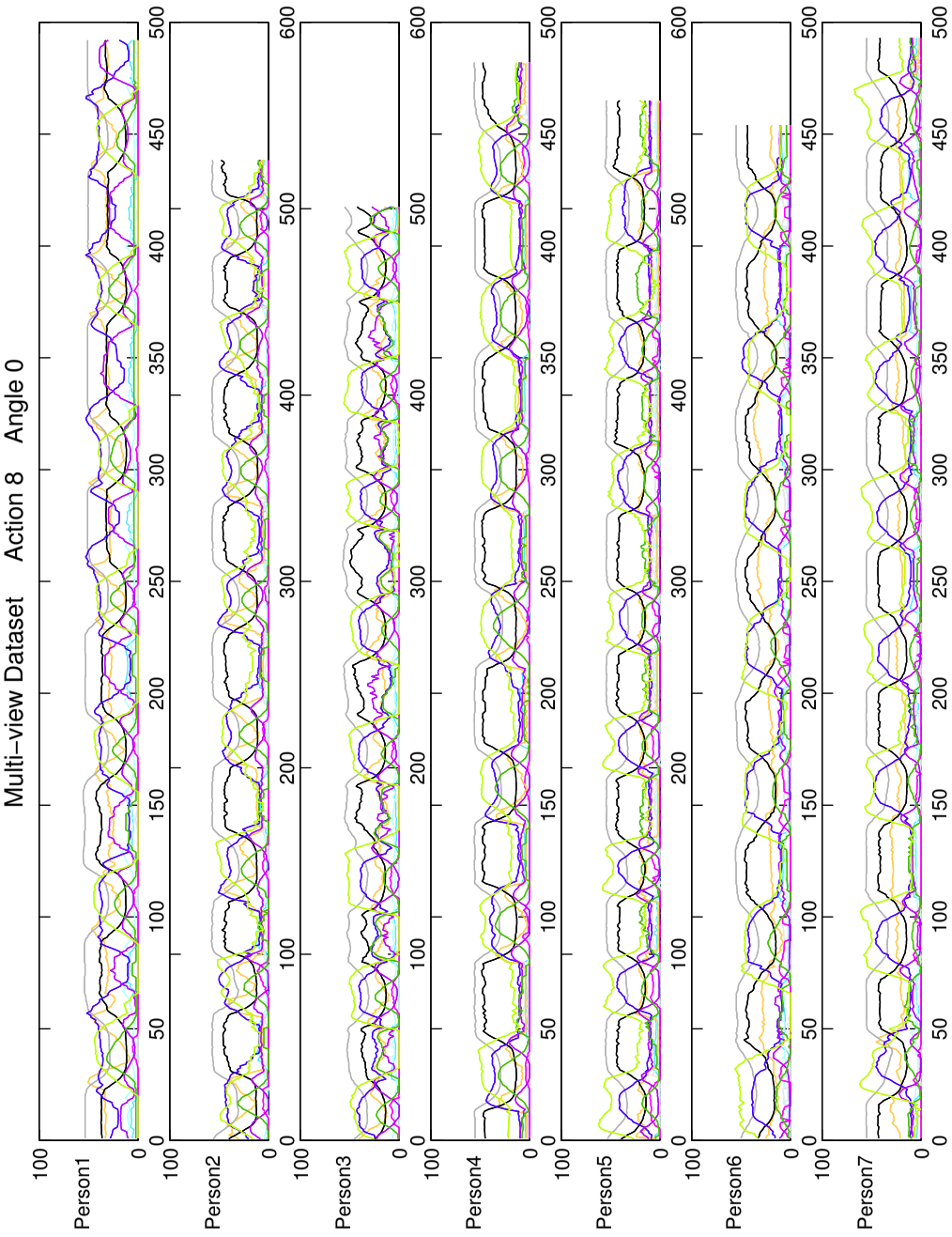

Figure C.11: The feature traces of Action 8 before processing.

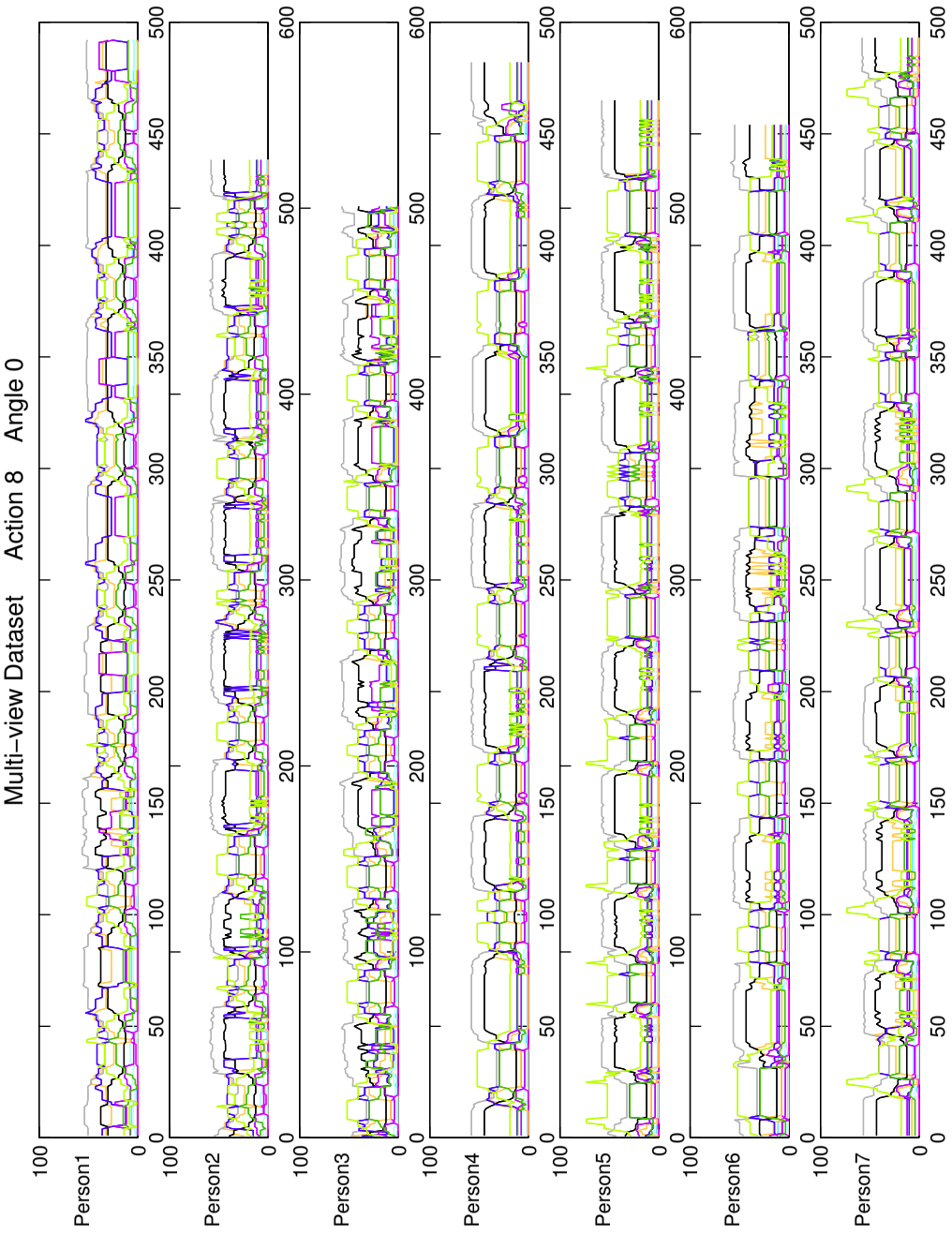

Figure C.12: The Action 8 traces after clustering

Appendix D

CD: Dynamic action sequences

The accompanying CD contains many animated sequences relating to the dissertation. The contents of the disk is as follows:

1. A visual tour of negative space: A number of dynamic images illustrate the ideas behind negative space analysis
2. Database 1: Animated colour coded negative space sequences.
3. Database 2: Animated sequences from the multi-view database