A HMM Framework for Motion based parsing for video from Observational Psychology

Daire Lennon, Naomi Harte & Anil Kokaram
Electronic Engineering Dept.,
University of Dublin,
Trinity College,
Dublin, Ireland;
lennondh@tcd.ie

Erika Doyle & Ray Fuller
Dept. of Psychology,
University of Dublin,
Trinity College,
Dublin, Ireland;
edoyle4@tcd.ie

Abstract

In Psychology it is common to conduct studies involving the observation of humans undertaking some task. The sessions are typically recorded on video and used for subjective visual analysis. The subjective analysis is tedious and time consuming, not only because much useless video material is recorded but also because subjective measures of human behaviour are not necessarily repeatable. This paper presents a HMM framework for content based video analysis to facilitate the automated parsing of video from one such study involving Dyslexia. The framework relies on implicit measures of human motion that can be generalised to other applications in the domain of human observation.

Keywords: HMM, Motion Based Parsing, Rotation Estimation

1 Introduction

Visual content analysis technology applied to surveillance applications is well established [1, 2]. Less explored is a class of human observational applications related to scientific study. The field of Psychology contains many such applications which may admit solutions from the visual content analysis domain. In surveillance applications, the behaviour of the people involved is not strongly constrained. This poses a challenge to understanding in that context. In scientific applications, it is likely that the behaviour being observed is restricted in some way in order to make measurements or other inferences. In Psychology this may be in a controlled environment or a particular set of movement/behaviours are being observed in a natural setting. This implies that content analysis could bridge the semantic gap more easily in that domain.

Previous work in Psychology has asserted a connection between reflex movement in children and a Specific Learning Disability (Dyslexia) [7]. To investigate this connection Psychologists at Trinity College `www.dysvideo.org` have designed a set of experiments that attempt to quantify this relationship in a study based on 150 children. The central idea is that children with a propensity to develop Dyslexia are unable to execute particular movements without some unavoidable associated reflex. Figure 1 shows an example of one such movement. The experimenter rotates the head of a child right and left. While doing so any involuntary bend in the arms is noted. The idea is that the presence of that reflex is in some way correlated with the presence of Dyslexia.

Work presented in Joyeux et al[3] reports on the **DysVideo** project at Trinity College that was set up to observe the development of 150 children between ages 4-7 years. Video recordings are made of children observed though 3 sessions, each of 20 mins duration, and 6 months apart. Unfortunately, in each recording of 20 mins, less than 5 mins is useful material. Children may take a long time to settle down and may need to be cajoled through each session. The

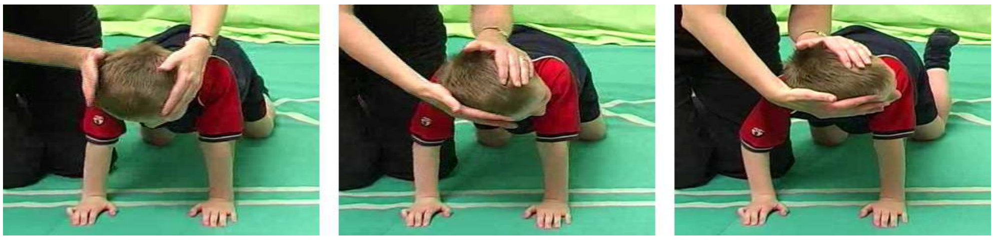


Figure 1: A demonstration of the ATNR exercise

Dysvideo project exploits automated content based audio and video analysis to allow Psychologists to index directly the useful portion of the video. Preliminary work on automated parsing was presented in Joyeux et al[3]. The focus there was the framework design and algorithms for coarse parsing and encoding of metadata by exploiting the audio stream. The parsing provided was good enough such that the estimated index points were guaranteed to contain the useful visual information. The size of the indexed portion typically contained about 20 seconds of material before the start and after the end of the actual useful motion experiment itself. This was adequate for the psychologists to browse quickly to the start of the useful recordings and perform the subjective motion measurement.

Recall that the point of the programme is to measure the presence of certain motion based reflexes in children. Currently for psychologists the only reliable way of doing this is for a human to assess the degree to which the child cannot hold a particular position. It is true that motion tracking equipment exists for this purpose, but magnetic based tracking is expensive, while visual marker based tracking would require the cooperation of the child in wearing the markered suit and not damaging the markers for future use. Quantitative visual motion measures can be designed by estimating the motion of the particular limb in the field of view. The important observation here is that this is only possible by ensuring that the motion measurement of the correct limb portion is being made and starts at the right time. Therefore for quantitative assessment of motion a finer granularity of parsing is needed. Work in this context was presented recently by Kokaram et al [4]. There the idea was to use an explicit measure of rotational motion to index measurable episodes directly. This work proposes instead a more robust framework for inference based on motion using the Hidden Markov model. The idea is to train HMMs to detect the specific type of motion required. This work also improves upon the range of motion features used and proposes the use of a novel feature for parsing: the curl of the estimated motion field.

The next two sections provide background for the reader that is useful in appreciating the context of the parsing algorithms developed. Section 5 the introduces the new features and the remaining sections present the use of the HMM framework.

2 Overview

Fig 2 shows the difference between the markers provided by audio parsing [3] and the exact start and end of a particular session, Test 10. The audio parsing is achieved by allowing the user to insert a specific audio tone into the recording using a handheld PC (PalmPilot). Postprocessing the audio signal [3] allows DTMF audio tones to be detected . The figure illustrates the importance and reliability of using audio markers to reject unwanted content. In this Test, the experimenter firmly rotates the head of the child while the child is on all fours. The hypothesis is that in children with a *retained reflex* one of the arms will bend at the elbow involuntarily during the motion of the head. To isolate the relevant content for analysis it is necessary to i) locate arms during that motion and ii) identify video portions during which the head is rotated. For some time before the start of the actual experiment, the child is coached and may undergo a few trials, in addition the child may move around and simply not be in the field of view. During the relevant video portions, the arm location will be more stable, and the rotation of the head more coherent. Therefore these features can be used to index the video with increased granularity. The starting point is an estimate of body location.

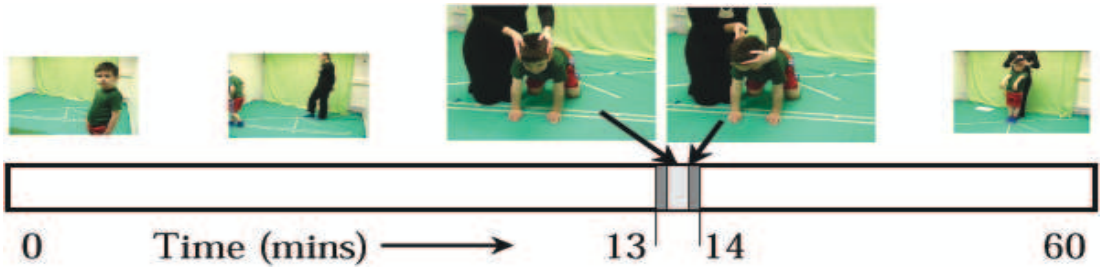


Figure 2: An example of content throughout a typical recording of 60 mins. The location of the audio markers coarsely delineates the relevant video for one Test. This is shown at mins 13 and 14. The actual useful content lies inside this period, indicated by two typical frames. The paper focuses on using motion based features for delineating this exact start and end of the experiment within the coarse audio marker period. Note that the idea of coarse indexing is to quickly reject the outlier content e.g. when the child is not in view, not cooperating, or not assuming the start position.

3 Body Localisation

Head and arm localisation is facilitated by skin detection. This is achieved by a simple colour segmentation process. The requirement is to configure a label field $l(\mathbf{x})$ that is 1 at pixel sites $\mathbf{x}$ containing skin and 0 otherwise. The algorithm is as follows.

1. Candidate pixels expressing skin ($l(\mathbf{x}) = 1$) are detected by colour thresholding (from [8]) using the following criterion

$$l(x) = \begin{cases} 1 \quad \text{if} \begin{cases} (R > 95)\&(G > 80)\&(B > 40) \\ \&(R > G - 15)\&(R > B) \\ \&((R - min(G, B)) > 10) \end{cases} \\ 0 \quad \text{Otherwise} \end{cases} \tag{1}$$

   The various parameters used in delineating the colour region were determined from the lighting used in the pictures recorded. This is the same throughout 100 hours of recording. The first two criterion delineate skin colour, while the last one rejects false alarms due to pixels that are near grey or near yellow.

2. The label field $l(\mathbf{x})$ is post-processed to smooth the surface. This is achieved using morphological closing with a dilation element of 3 pixels and a erosion element of 4 pixels.

As shown in figure 1, the arms are generally the largest area of skin exposed in the view. In addition they are near vertical. Hence a vertical sum (integration) of the label field yields modes corresponding to the horizontal position of the arms. Given the detection field $l(\mathbf{x})$ the vertical projection is defined as $p^v[h] = \sum_k l(h, k)$. Noise in $p^v[h]$ is removed by filtering with a Gaussian filter with 9 taps and variance 1.5. To detect modes in $p^v[h]$ the two most significant maxima are selected that are at least 50 pixels apart in PAL videos. This allows robustness to false alarms within a single arm segment. Figure 3 clearly shows the correlation between lobes and horizontal arm location for 2 different recordings of 2 different children. Note the false alarms due to poorly detected skin in the background (due to strangely coloured walls) are rejected with this process.

Locating the hands is achieved through the horizontal projection of the label field $p^h[k] = \sum_h l(h, k)$. The first maxima corresponds roughly to the middle of the hand position because of the orientation of the child in the view. This is shown in Figure 3. The very first non-zero projection corresponds to the start of the hand location. The hand size is estimated to be the width of the lobes corresponding to each arm, D. The wrist location is hence taken to be $1.5D$.

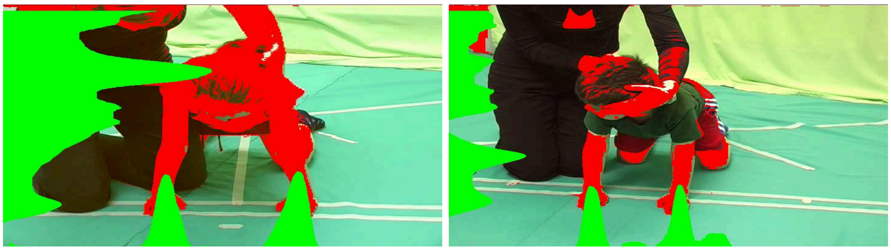


Figure 3: Example frames from different sequences showing the results of skin detection and hence body localisation. The detected skin pixels are coloured in red. The horizontal and vertical projections of the label field are shown in green along the left and bottom edges of each frame. This illustrates that the lobes in vertical projection correspond to arm location. The first mode in vertical projection corresponds to arm location.

In addition, the average forearm length is approximately 2.5 times the hand width in this view, hence the location of the elbow can be roughly delineated vertically. This enables a bounding box to be placed that contains the hand and arm locations. The process is found to be better than 99% accurate in these sequences, provided the child adopts the correct position. Typical results are shown in Figure 3. For video examples, see www.sigmedia.tv/research/indexing/dyslexia/.

The location of the arms is used to bound the head location horizontally. Therefore, head location is assumed to be contained within a column of the image bounded by the left and right arm locations. Unfortunately, detection of the head using projections is not reliable because in horizontal projection the face and arms of the experimenter can often cause ambiguity.

4 Motion Based Parsing: Features

The ultimate aim is to detect the contiguous sequence of frames showing rotation of the head. Given the delineation of a region containing the head, it is possible to estimate directly that rotation, and that can be used to attempt parsing as in [4].

For each frame, gradient based motion estimation [5] was performed where the previous frame and motion vectors were stored. Motion vectors either side of the located arms were removed for improvement in the information about the child since only those vectors were applicable. Using the motion vectors themselves and plotting there perpendicular lines in an accumulator array resulted in finding an approximate centre of rotation. This is based on the principle that a perpendicular to a tangent of a circle will always pass through the centre of the circle. The central four images in figure 4 shows a selection of accumulator arrays ranging over a head rotation sequence. The distance between the centre's of rotation from frame to frame was consistently small and stable when rotation was occurring. This was used as a feature to indicate rotation.

In this work, the use of the curl of the motion vector field is also exploited to yield an implicit measure of rotation. Given a motion vector at site $\mathbf{x}$ is specified as $\mathbf{d} = [d_1(\mathbf{x}), d_2(\mathbf{x})]$, where the horizontal and vertical displacements are d_1 and d_2 , the curl of the motion at that site $\mathcal{C}$ is given as below

$$\begin{aligned} \mathcal{C}(d_1, d_2, \mathbf{x}) &= \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ \frac{d}{dx} & \frac{d}{dy} & 0 \\ d_1 & d_2 & 0 \end{vmatrix} \\ &= \vec{k}(\frac{d(d_1)}{dy} + \frac{d(d_2)}{dx}) \end{aligned} \tag{2}$$

The curl therefore is a vector pointing out of the image plane with a length that is proportional to the amount of rotation at that site. The bottom four images in figure 4 show a selection of the curl matrices ranging over a head rotation sequence.

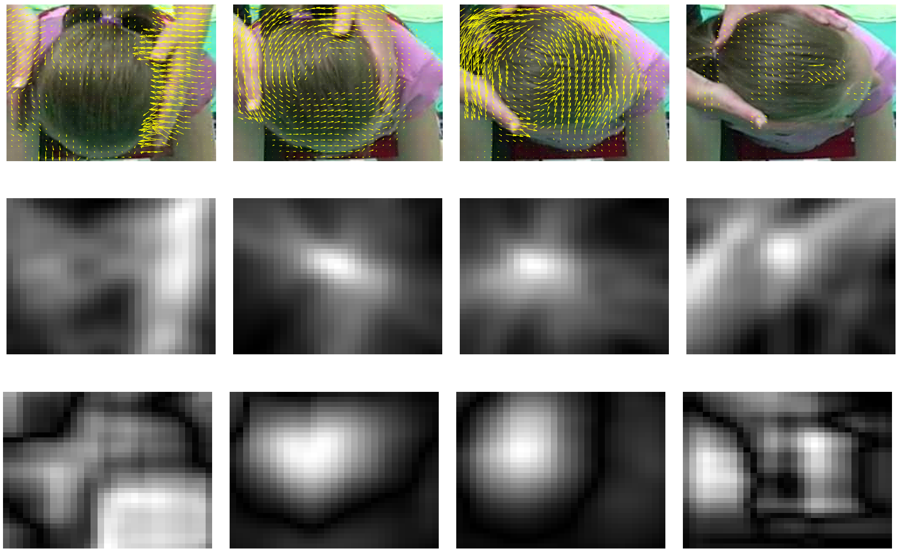


Figure 4: The top four images show a selection of frames used to demonstrate a sequence of head rotation. The central four images show the same sequence for the accumulator array and the bottom four images show the sequence for the curl matrix. All of the above images have been zoomed in on to improve clarity.

As can be seen in the central images in figure 4, the peaks in the curl correspond well to centre of rotation and there is stability in position during rotation. The distance between the maximum peaks from frame to frame indicates the level of stability and it can be observed that there is relatively little motion of the maximum peak during rotation. The distance measure between frames for the maxima was used as a feature. It is also observable that the peak value of the curl rises and falls relatively smoothly with rotation. The derivative of the graph of peak values was also used as a feature.

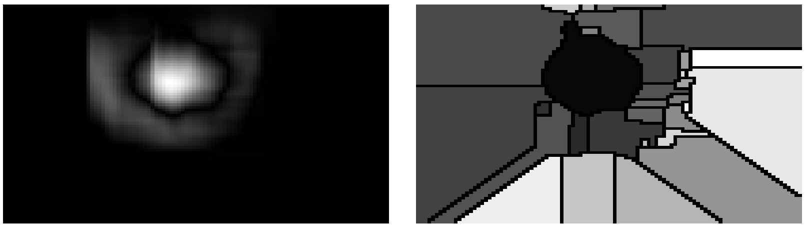


Figure 5: Example frame used to demonstrate watershed segmentation of the curl surface.

The main mass of the maximum peak during rotation are observed to be reasonably constant. To use this information, it is necessary to segment out the main mass of the curl surface. This is done with watershed segmentation [6] on that inverted curl surface, as demonstrated in figure 5. The masses of the maximum peak for each frame was used as a feature in our HMM models.

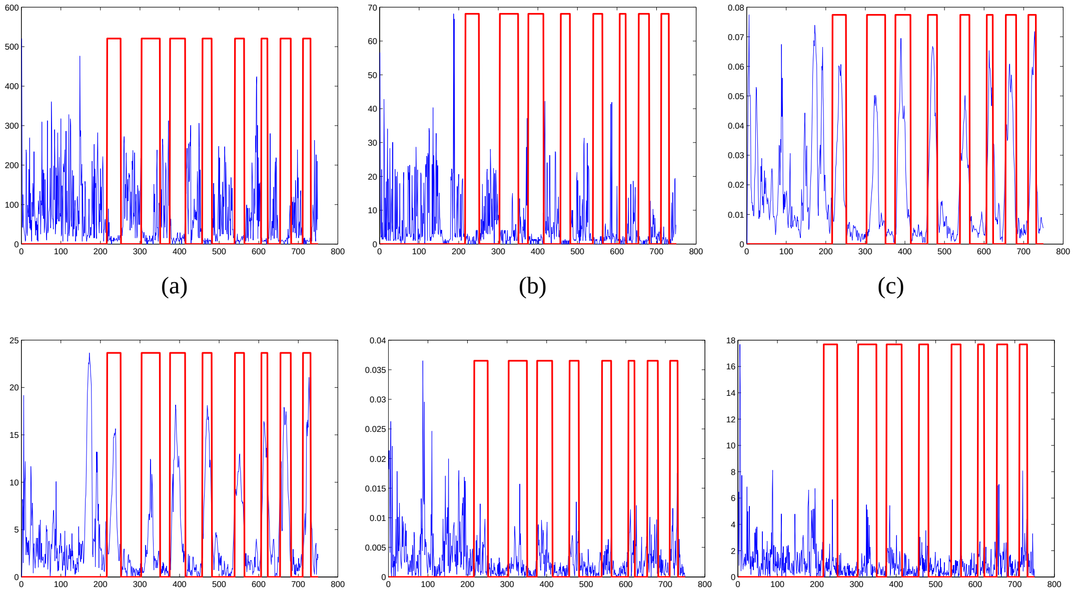
It was also found that the graph of the peak masses was consistently rising and falling during rotation. The derivative of the graph of the area under the maximum peak for the entire sequence was also considered to contain useful information and was so used as a feature vector.

The set of features used for motion based parsing can be summarised as follows:

1. The distance from maximum peak to maximum peak of the accumulator array from frame to frame. See figure 4 for illustration of process.

2. The distance from maximum peak to maximum peak of the curl surface from frame to frame.

3. The value of the maximum peak of the curl surface.



(a) (b) (c)

(d) (e) (f)

Figure 6: An example set of feature vectors to be used in the training of an HMM. The red marking represent manual segmentation markers. Figure (a) & (b) represents the distance between the maximum peaks in the accumulator array and curl matrix from frame to frame respectively. Figure (c) is the graph of the maximum peak values in the curl matrix and figure (d) is area under the peak surface. Figure (e) & (f) are the derivatives of the graphs of figure (c) & (d) respectively.

4. The area under the maximum peak surface, as segmented by the watershed algorithm on the curl surface.

5. The derivative of the graph of the maximum peaks of curl surface for the entire sequence.

6. The derivative of the graph of the area for the maximum peaks of curl surface for the entire sequence.

This feature vector $\mathbf{f}_n$, containing these 6 measures, is measured for each frame in the sequence. Figure 6 shows the evolution of each component over a an entire exercise sequence.

5 Motion Based Parsing: Modelling

The HMM is a well established framework for time series modelling and has been very successful in speech recognition [9]. The idea here is to use two HMMs to model the evolution of the multidimensional feature vector $\mathbf{f}_n$. One HMM models the non-rotation segments and the other models the rotation segments. Given a manually labelled training set, the parameters of each HMM can be established. These models are used in parsing new examples. The use of HMMs for modelling video features is a relatively recent idea, exploited successful for sports by Rea et al [10].

Fig 5 shows the structure of the HMM used in this framework. It is a four state HMM with the ability to transit from any state to any other state. Note the difference from the standard left to right models employed in speech recognition applications. The HTK Speech Recognition Toolkit was used to initialise and train the statistical parameters of our HMM models. Gaussian distributions were assumed with single mixtures per state distribution. A total of 23 videos were selected from session 1, 16 to be used for training and 7 for testing. The criterion used to select the videos was based on arm separations. An example of training data is shown in in figure 6. The performance of the resulting models is presented in section 6.

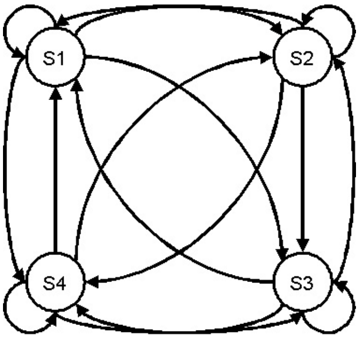


Figure 7: A 4 state HMM Model with possible transitions in all directions. This was the model used for both the rotational and non-rotational HMM's. Although with differing transition probabilities and feature statistics.

6 Results

Video selection was difficult because the quality of the child's motion was essential. The main goals of the video recordings were the ability for direct analysis by the psychologists, secondary was our analysis of the videos. As such the demonstrator focused more on just recording the exercise for human evaluation as opposed to improved conditions required for our evaluation. So if the size of the child is too small relative to the frame the motion of child will be too small relative to the larger motion of the experimenter which would be clearly visible in a shot of wide angle. Arm separations are a clear indicator as to the size of the child relative to the frame size i.e. zoom factor. The selection of 23 videos were chosen to best represent the exercise on the bases that they were the videos with the largest visible children.

The 23 videos had to be manually segmented. The frame numbers for the start and end of any rotational occurrence were noted for each sequence. These manual segmentations were compared with the outputs generated by the HMM models. The analysis of the manual segmentation markers versus the HMM markers had to take into account the human error. It can seen for individual video comparisons that there is a difference between the two markers of a couple of frames. Human observations follow the exercise start and finish more so than the rotational occurrences. Taking into account that human observations are more subjective, an error of 12 frames was deemed acceptable. The results below reflect this adaptation.

$$Recall = \frac{Correct}{Correct + Missed} \qquad Precision = \frac{Correct}{Correct + False}$$

	Recall	Precision
Test Video 1	81.7109	86.8339
Test Video 2	77.7778	82.3529
Test Video 3	47.2603	50.4263
Test Video 4	62.7525	64.4617
Test Video 5	55.9361	77.044
Test Video 6	62.6935	68.1818
Test Video 7	63.6879	72.1865

Results to date have been extremely promising. By modelling the described features with the HMM framework the detection of rotational occurrences has been improved significantly over data thresholding, which was the previous method of data analysis. The performance of test video 3 is poor and may be attributable to poor motion vectors from the original data. This will be further investigated.

7 Final Comments

Future work includes trying to gain a better understanding of the contribution of the features to the recognition framework. Refining the manual segmentations to only take into account actual rotations and not other motion. Adapting the algorithm to poorer quality videos, i.e. ones where the child appears small in the frame. These refinements will hopefully improve the HMM Models and accuracy of detection.

References

[1] Edward Y. Chang and Yuan-Fang Wang. Video surveillance. In *First ACM SIGMM international workshop on Video surveillance*, November 2003.

[2] Arun Hampapur. S3-r1: The ibm smart surveillance system-release 1, June 2005.

[3] L. Joyeux, E. Doyle, H. Denman, A. C. Crawford, A. Bousseau, A.Kokaram, and R. Fuller. Content based access for a massive database of human observation video. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 45–62, October 15-16 2004.

[4] A. Kokaram, E. Doyle, D. Lennon, L. Joyeux, and R. Fuller. Motion based parsing for video from observational psychology. In *Proc. SPIE Vol. 6073, p. 265-274, Multimedia Content Analysis, Management, and Retrieval*, Jan 2006.

[5] A. C. Kokaram. *Motion Picture Restoration: Digital Algorithms for Artefact Suppression in Degraded Motion Picture Film and Video*. Springer Verlag, ISBN 3-540-76040-7, 1998.

[6] Vincent Luc and Pierre Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 13:583–598, June 1991.

[7] M. McPhillips, P.G. Hepper, and G. Mulhern. Effects of replicating primary-reflex movements on specific reading difficulties in children; a randomised double-blind, controlled trial. *Lancet*, 355:537–541, 2000.

[8] Peter Peer, Jure Kovac, and Franc Solina. Human skin colour clustering for face detection. In *EUROCON 2003 - International Conference on Computer as a Tool*, 2003.

[9] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[10] Niall Rea, Rozenn Dahyot, and Anil Kokaram. Modelling high level structures in sports with motion driven hmms. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2004.