Action recognition in multimedia streams

Rozenn Dahyot, François Pitié, Daire Lennon, Naomi Harte, and Anil Kokaram

Trinity College Dublin Rozenn.Dahyot@cs.tcd.ie.

It is well accepted that the rise in the proliferation of inexpensive digital media collection and manipulation devices has motivated the need to access this data by content rather than by keywords. The requirements of content based access are well understood by the digital media research community and there is no need to elaborate further here. Parsing multimedia streams by detection and classification of action implies modeling the dynamic nature of visual and audio features as they evolve in time. The Hidden Markov Model (HMM) has long been used to model dynamic behaviour in audio signals. Its power to capture complex behaviour in that domain has led to widespread use in visual content analysis because of the non-stationarity inherenet in those signals. However, subtleties in the application of HMMs are often unclear in the use of the framework in the visual processing community and the latter portion of this chapter sets out to expose some of these. Three applications are considered to motivate the discussions: actions in sports, observational psychology and illicit video content.

Sports: Work in sports media analysis and understanding has been conducted for a decade now with clear motivation provided by the huge amount of sports media broadcasting on internet and digital television. An overview of content analysis for sports footage in general can be found in [22]. Action recognition here involves detection of certain *plays* and *situations* as dictated by the game domain e.g. pots, goals, wickets and aces.

Illicit Content: The distribution of pornographic materials has also benefited from the digital revolution [6]. This kind of material is illegal in the workplace and is referred to as *illicit content* in this article. The issue of filtering this material has been of major concern since the introduction of the web in the early 1990's. Pixalert's 'Auditor' and 'Monitor'[1], FutureSoft's 'DynaComm i:scan' [2] and Hyperdyne Software's 'Snitch'[3] all provide image and

[1] http://www.pixalert.com/product/product.htm

[2] http://www.futuresoft.com/documentation/dciscan/imagerecognition.pdf

[3] http://www.hyperdynesoftware.com/clean-porn.html

text filtering for remote scanning of e-mail, hard disks and peripheral storage devices (e.g. USB memory keys). While there has been noteworthy activity in research into content-based analysis of illicit images [17, 19, 4, 39, 3, 2], there has been little work in spotting illicit activity in video streams. The need for such work has become stronger with the popularity of media sharing (via YouTube and Video Google for instance) and the requirement for host sites to police usage. Action recognition in this context requires multimodal analysis of motion and audio features.

Scientific: Observation of people occupies much of the time of the behavioural psychologist. The digital revolution has allowed video to be recorded easily enough so that behvioural assessments are in principle more scientifically recorded and analysed. In the experiment discussed in this paper, over 300 hours of video were recorded of children undertaking specific movement therapies. Reviewing and scoring the video of each subject is therefore an arduous task made difficult by the lack of easy indexing to the key actions of interest. Action recognition in this context involves the detection and parsing of video showing rotational motion in the region of the subject's head (see Fig. 8). This example illustrates a little known use of HMMs i.e. not only to classify temporal activity, but also to parse a sequence according to that activity.

Broadly speaking there are two approaches to parsing through action. In certain cases (*Direct Parsing*), specific features can be directly connected to the action of interest and a relatively *thin* inference layer then yields decisions and hence a parsed stream. In other situations (*Model Based Parsing*), the connection between features and actions is not straighforward and a *heavier* inference layer is needed to articulate the feature information in order to yield a decision. In all cases, motion of objects or the camera itself is important for action parsing, and so motion estimation and object tracking are key tools in the content analysis arsenal. In broadcast footage, where the editing itself is an indication of action, preliminary shot cut detection allows visual material in each shot to be analysed in separate units. In scientific or surveillance type footage the actions of interest occur as impulsive events in a continuously changing stream of material.

1 Direct Parsing for Actions

Both sports analysis and illicit content identification contain good material for discussing direct parsing. When features are strong enough to yield detection directly, a useful pre-processing step is the delineation of media portions which are most likely to contain that action. In illicit content analysis, the presence of large amounts of skin coloured regions are a strong indicator of video clips of interest. Skin regions occupy a relatively narrow range in the colour spectrum and Dahyot et al [35] compute the posterior probabilitity $p(\text{skin}|z)$ that each pixel z belongs to the skin class. This p.d.f. is obtained empirically using skin and non-skin reference histograms from the open-source filtering Poseia project[4]. While this formulation treats pixels independently, it is a sufficient model for the initial skin segmentation. A skin binary map is then generated by thresholding the probability map.

Sport videos usually show a finite number of different views and the actions of interest are only contained in a subset of views. View classification can be achieved in sports with either low level, direct feature manipulation or model based recognition. Since the principal actions usually take place in views that contain mostly the playing area, and the playing area is usually of a predefined high contrast colour, colour features from each frame allow quick identification of the shots that contain player action. This is a well established idea used to good effect by several early authors [12, 18, 20, 5, 15]. Figure 1) shows example frame segmentations using colour thresholding of the average frame colour used to good effect in [11]. The playing area segmentation implicit in this shot segmentation exercise then yields the geometry of the view, and the delineation of the playing area itself within the view. The Hough Transform is typically used to do this [12, 20]. See figure 1 for an example.

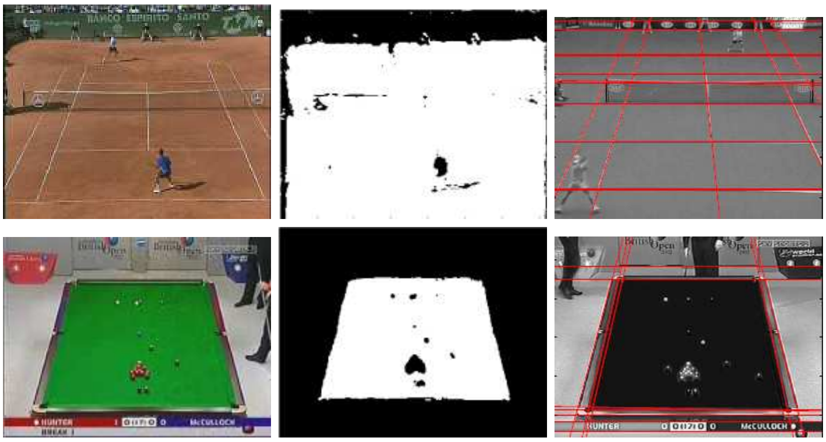


Fig. 1. Top row: Tennis frame showing unsupervised segmentation of the playing areas using colour information, and calibration of the playing area (far right). Bottom row: The same information for snooker.

1.1 The actions

Having delineated the important video material and the active area in the frames, motion or change analysis can directly be matched to certain actions.

[4] http://www.poesia-filter.org/

For instance, Denman [12] observed that the position of the pots in the snooker table were fixed in the relevant view, and the location of the pots could be determined accurately in the calibration stage. Hence colour histogram change analysis in the region around each pot could detect a ball *pot* action event. Dahyot et al [10] observed that racket hits in tennis and bat hits in cricket are unique impulsive sounds in the audio stream. Principal Component Analysis (PCA) from the audio tracks associated with relevant views, can be used to design specific filters (thresholding of the PCA feature distance from the training cluster) to perform detection of these sounds to near 100% accuracy. As the sound is associated with a specific dynamic action, this means that the action can be detected with high reliability, in effect by thresholding a single PCA-derived feature.

Motion analysis of course yields a much richer action detection process. For instance, although collision of snooker balls can be heard through the audio track, the strength of that sound is not significantly higher than the background noise and snooker ball collision through audio alone is unsuccessful. Both global/camera motion and local object motion yield information rich features. Global motion estimation (6 parameter affine motion) can be achieved with weighted least squared methods e.g. [31, 13, 7]. Kokaram et al [21] shows that global motion can be connected directly to *bowler run* up and *offside/onside shot* actions. This is because in cricket broadcasting the camera zooms in as the bowler runs into throw the ball, and then zooms out and pans left or right to follow the ball after it is hit. The rough run of play action in soccer can also be characterised by the global translation of the camera move [28].

Local motion information contains the motion of the players and sport objects and hence is directly relevant to the play. Typically the objects of interest are first segmented from the playing area in the field of view and then tracking is instantiated in some way. Both Ekin [14] and Rea et al [34] exploit schemes based on colour histograms. However, Rea et al adopt the popular (at the time) particle filter tracking approach while Ekin adopted a determinsitic matching scheme that selected the matching tiles on a fixed grid over the plkaying area which contained the object in question. Rea et al also introduced the notion that, given the calibrated view provided from Denman et al [12], it is possible to alter the size of the bounding box containing the object to be tracked so that it compensates for the view geometry. This is quite an important idea for sport action tracking where the view geometry will affect the size of the object and hence the ability to match any template colour histogram. Nevertheless, Pitié et al [32] point out that colour based segmentation in sport is able to remove much of the ambiguity inherent in many *hard* tracking problems. In other words, the regions of the playing area that are not part of the playing area colour, are likely to be positions of objects to be tracked. This idea leads to a viterbi scheme for tracking that selects the best path through candidate “blobs” of interest in each frame of the sequence. This latter idea is much more computationally efficient and robust than particle filters in the sport application.

Given motion trajectories of objects it is possible to directly classify object actions in some applications. For instance, in snooker loss of tracking “lock” near a pot in the table indicates that a ball has been potted. Loss of lock can be established by thresholding the likelihood energy of the tracker in each frame for each object [33]. In that work, a ball collision is detected by identifying changes in the the ratio between the current white ball velocity and the average previous velocity. If the ball is in the vicinity of the cushion, a cushion bounce is inferred. Given that the physics of colliding bodies implies that at collision, changes in velocity in one direction are typically larger than another, a change in velocity of 50% is used to indicate of a collision. A flush collision is inferred when velocity changes in 50% in both directions.

1.2 Exploiting the Motion Field

In illicit content analysis the situation demands a more implicit motion feature extraction approach. The problem is that only a portion of the skin covered regions would yield information amenable to further analysis and it is not possible to easily further delineate any obvious feature for tracking on the basis of colour or texture alonw. Instead, local motion over the entire detected skin area can be used as a feature to segment objects or regions for further analysis. Using motion extracted from the MPEG compressed stream leads to a computationally efficient procedure.

In order to segment the local motion regions, global motion must be compensated for. Macroblocks that contain less than 30% skin pixels are cited as non-skin blocks and are used to estimate this motion. The blocks containing low texture (with low DCT coefficient energy) are removed from further analysis as they will contain unreliable motion information. The mode of the 2D motion histogram of these motion vectors yields an estimate for global motion. Segmentation using the raw MPEG vectors is likely to lead to temporally inconsistent masks because MPEG motion, based on block matching is likely to be temporally poor. To alleviate this somewhat, the motion field is filtered with a 3D vector median opreation using the ML3D filter outlined in Alp et al. [1]. Once the vectors have been compensated for global motion, they are clustered using K-means, assuming only two clusters are required for foreground/background. K-means is used since it is a computationally efficient clustering algorithm and gives satisfactory results compared to the watershed segmentation used by [8]. The region of interest is then the logical ’and’ of the skin map and this foreground motion map.

Figure 2 shows the binary skin image and the motion compensated segmentation with overlaid motion vectors for a still from *When Harry met Sally*. Using the motion information helps to segment relevant skin region with higher accuracy. Detecting periodic motion behaviour has become increasingly popular for retrieval in video [9, 29, 16]. The motion estimated here can be directly associated with periodicty of that skin region and thus a notion of *illicit video* [35].

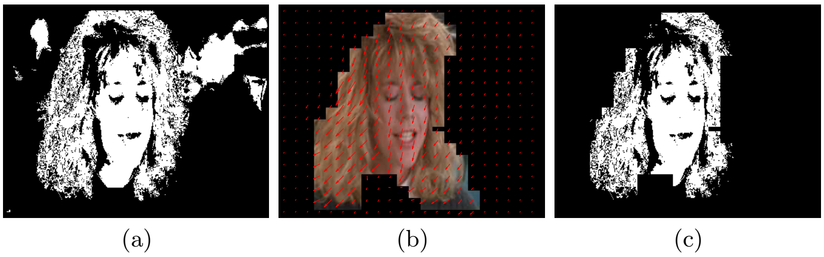


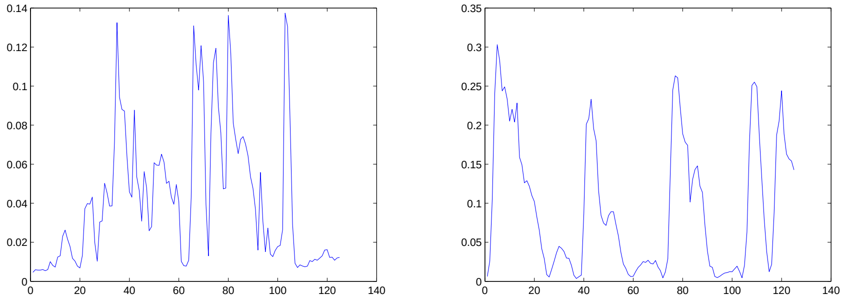
Fig. 2. (a) Binary map of the skin segmentation; (b) Motion segmentation with overlaid motion vectors; (c) Binary 'and' of motion and skin segmentations

1.3 Exploiting audio

Even when not watching the video content from a multimedia stream, the nature of the stream can still be understood from the audio information alone. Examples of applications can be found in sport video indexing as discussed above. This is true also of pornographic content. Periodic audio signals can be indicative of illicit contet. The famous scene from the movie *When Harry met Sally* (Sally's simulation of an orgasm, which is a series of moans and screams) sevres to illustrate the point. The scene starts with a conversation between Sally and Harry. The loudness of the audio signal is computed over non-overlapping temporal windows of $0.04s$ (duration of a 25fps video frame). For analysis of periodic patterns, a 5 second period is used corresponding to 125 measurements of volume. Figure 3 presents two 5 second periods and confirms that a periodic pattern is exhibited during the *illicit* extract (b) more so than during the conversation (a). Periodicity in the signal is usually analysed by autocorrelation, circular correlation or periodogram [38, 36]. Autocorrelation is used here and the autocorrelation for the two signals in figure 3 is given in figure 4. Peaks appearing in (b) show that the signal is periodic.

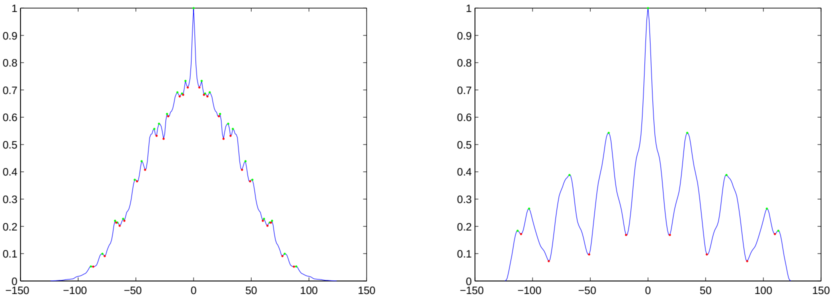
The key is to define a measure to discriminate autocorrelations of classes similar to (a) and (b) (cf. figure 4). A simple measure is to compute the difference between the surface defined by the minimas and the maximas of the autocorrelation. This is illustrated in figure 5 for the same audio extracts (a) and (b).

Figure 6 shows this periodicity measure during the whole scene of *When Harry met Sally*. The measure is low at the start as only a conversation occurs between the two main characters. Then starting at 95 seconds, the periodic pattern begins. In this case, periodic moaning and screaming appears on the audio data. By the end of the scene, standard conversation takes place again and the measure of periodicity decreases.



(a) (b)

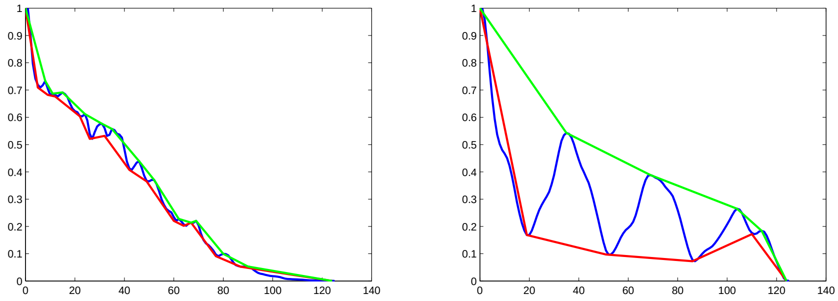
Fig. 3. Audio energy computed over 5s when Sally talks to Harry (a), and when Sally is simulating (b).



(a) (b)

Fig. 4. Autocorrelation of the energy in the audio data with their maxima (green dots) and minima (red dots).

Using a threshold of 4 to detect illicit content when the measure exceeds this value, leads to a usable action spotting algorithm. It performs a perfect segmentation in the scene of *When Harry met Sally* (cf. figure 6). The method has been assessed first on non-illicit materials (20 minutes of extracts from movies and music videos) to evaluate the false alarm rate of the method. Various audio sources was used (music, speech, explosion, scream etc.), and in all those, the false alarm rate is rather low at 2%. The detection rate is more difficult to assess as periodic sounds do not occur all the time in the audio stream. Ten minutes of eight different extracts of illicit materials showing periodic sounds have been used. Five extracts corresponding to 9 minutes of the test have been properly detected. Three short extracts (representing 1 minute of recording) are missed. On those three files, a mixture of sounds is



(a) (b)

Fig. 5. The measure of periodicity on the half-autocorrelation is computed by the surface between the green curve (defined by the maxima's in figure 4) and the red curve (defined by the minima's in figure 4).

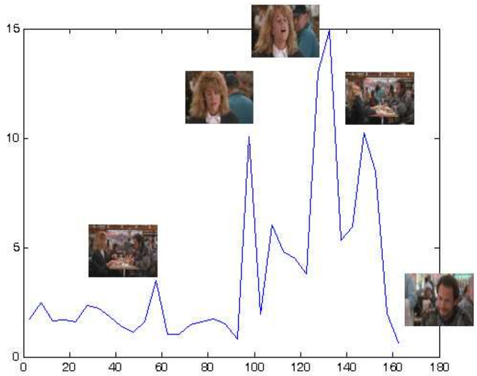


Fig. 6. Measure of periodicity in the scene of *When Harry met Sally* w.r.t the time (in seconds).

occurring (speech or music) masking the relevant periodicity on the loudness feature.

2 Model Based Parsing

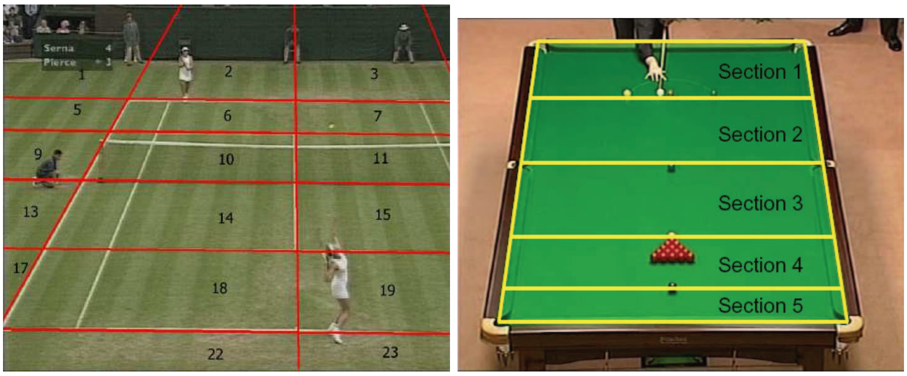
To gain deeper access to action sementics some form of inference layer is needed for processing the temporal evolution of the motion feature. The HMM has been heavily exploited for this purpose. Traditionally, HMMs are well established as a means of modelling the evolution in time of spectral features in the speech envelope. The underlying IID (Independent and Identically Distributed) assumption of HMMs for audio is that there is no correlation between successive speech vectors. That has strongly motivated the use of features such as cepstrum that themselves inherently capture the dynamic characteristics in speech. Feature vectors are generally augmented with first and second order derivatives to further improve speech recognition rates. The choice of features for visual applications is extremely diverse and in many cases ad-hoc. Visual HMM frameworks can be better designed by examining whether discrete or continuous density models are suitable for the application, whether feature sets are truly independent and hence full convariance models are not needed, and whether the HMMs are to be used for classification or recognition purposes. This is analagous to defining whether a speech recognition task is classification of isolated units, or full recognition where both unit classification and parsing are jointly performed. To understand how HMMs can be used for action classification, consider the two examples as follows.

2.1 Action in Sports

Given the extraction of the motion trajectories of objects explained previously, it is clear that the shape of that trajectory contains information about what is happening. A simple example is the trajectory of the white ball in snooker, if it traverses the whole table and comes to rest near a cushion, that is probably a conservative play. Trajectory classification then is very similar to handwriting recognition. Analogous to the approach used for on-line handwriting recognition [25], active regions are delinaeated in the tennis court and on the snooker table (fig. 7). Those regions represent the discrete states on which the trajectories of the balls in snooker and the players in tennis, are encoded. Hence as the ball and players move around on the playing surface they generate a time series of symbols.

Rea et al [34, 33] use a first order HMM to classify these sequences. By their nature, the sequences are discrete, and hence a discrete HMM is employed. A different model is trained using the Baum-Wesh algorithm. As the actions are well understood in terms of the geometrical layout of the table, the models can be trained using user inputs or training videos with ground truth. The types of actions amenable to analysis in this fashion are as follows.

Snooker	Tennis
Break building	Aces
Conservative play	Faults
Snooker escape	Double Faults
Shot to nothing	Serve and volleys
Open table	Rallies
Foul	



(a) (b)

Fig. 7. Spatial encoding of the playing area.

3 Action in Psychological Assessment

Action classification using HMMs in sport relies strongly on the pre-processing mechnisms and domain specific knowledge which allow that portion of the video containing the action to be pre-segmented for analysis. In the Dysvideo project (www.dysvideo.org) [24] the video recorded is of a single view in which a stream of actions are being performed continuously. Action recognition here involves the detection and parsing of video showing rotational motion in the region of the subject's head (see Fig. 8). What is required here is a process not only to identify the onset of the rotation exersise, but also to qualify when the head is rotating to the right or the left. This implies recognising the action and also using it to parse the video and it is possible to use the HMM here as well. This is subtle variation in the use of the HMM and here two **continuous** density HMMs are used - one representing rotation events, the other non-rotation events. Using classic Viterbi-based recognition, periods of rotation and non-rotation can automatically be distinguished [26].

3.1 Motion based features for human movement assessment

The rotation of the head is detected by analysing features of the motion flow in the video. To avoid dealing with the movements of the instructor, the analysis is restricted to the region around the head of the child. Head tracking is thus required, and a similar technique as previously discussed in this chapter has been implemented. A skin colour segmentation is first performed to isolate the child from the background. As part of the experiment, the child is required to wear T-shirt and shorts so a good part of visible skin belongs to the child. As shown in figure 9, the arms are well exposed in the view. In addition they are near vertical. Hence a vertical sum (integration) of the skin label field yields a

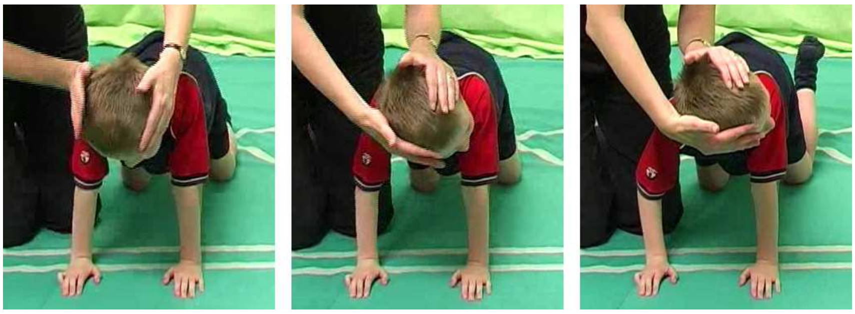


Fig. 8. A demonstration of the ATNR exercise

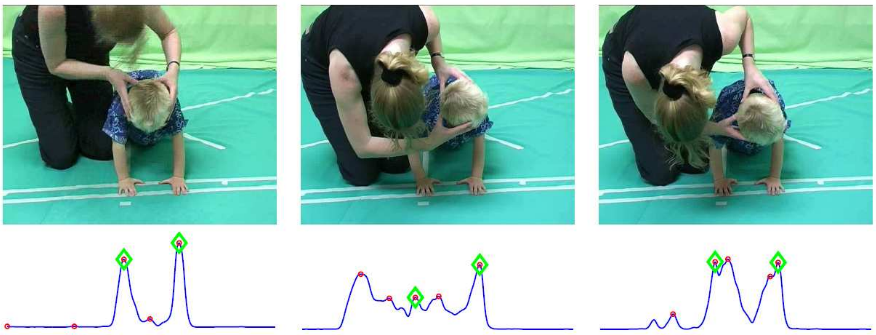


Fig. 9. Detection of the hands positions (green diamonds) of a child performing a psychological exercise [32]. The blue line shows the skin colour projection and the peaks give the candidate positions (red circles).

1D projection whose modes correspond to the horizontal position of the arms. The head position can then be found in between both arms.

As illustrated in Figure 9, occlusions by the instructor can create spurious peaks in the projection. To find the correct peaks, a Bayesian approach is adopted. At every frame, all the peaks of the 1D projection are collected as candidate positions. The ensemble of these candidate positions constitutes a trellis. The positions of the hands are retrieved by imposing some prior on the motion of the hands and running the Viterbi algorithm through this trellis to extract the most likely path.

With the child head isolated, features can now be derived to model the motion of the child. These features have to be capable of determining when the head of the child is rotating. Since rotation is a unique type of motion, gradient based motion estimation was performed [23] and the motion vectors for each frame were calculated for each exercise sequence. The calculated motion vectors are only capable of showing locally translational motion. However, looking at a larger scale, the spatial variations of the vector field can be used to identify non-translational motion. In particular, the rotational component of a vector field can be obtained by measuring the curl of the motion vector field. Denote as $u(x, y)$ and $v(x, y)$ the x and y components of the motion field between frames I_n and I_{n+1}. The locally translational motion equation at pixel (x, y) is given by:

$$I_{n+1}(x + u(x, y), y + v(x, y)) = I_n(x, y) \tag{1}$$

The corresponding amplitude of the curl for this 2D motion field is then defined as:

$$\mathcal{C}(x, y) = \frac{dv(x, y)}{dx} - \frac{du(x, y)}{dy} \tag{2}$$

The curl yields an implicit measure of rotation. Example of the curl field for an head rotation exercise is displayed on Figure 10. The main peak in the curl corresponds to the centre of rotation and its position remains stable during the rotation.

From the curl surface, it is possible to infer two essential features: the rotation centre and the size of the rotating object. The centre of rotation is given by the main peak in the curl. The estimation of the rotating object area requires to delineate the head with a watershed segmentation on the curl surface. The set of features is completed by adding the temporal derivative of the position and the size. The reasoning behind is that during rotation and non-rotation events, temporal variations of the object position and size are radically different. These four features are combined with two other features, which are described thoroughly in [27]. A total of six features is therefore used to characterise the rotation movement of the head.

3.2 Event recognition in psychological assessment

Using the feature set discussed, continuous density HMMs are trained and used in viterbi-based recognition to parse unseen video into periods of rotation and non-rotation. The rotation model $\mathcal{R}$ is associated with a dedicated continuous fully connected 4-state HMM. Other non-rotation events are modelled by another model $\overline{\mathcal{R}}$, which is also associated with a continuous fully connected 4-state HMM. For both HMMs, the likelihood of being in a particular state is defined by a single Gaussian distribution. Evaluating the MAP of a sequence of observations can be done using the Viterbi algorithm. To decide if a sequence is a rotation or non-rotation event, it is then sufficient to compute the MAP for each model and choose the most likely.

A naive approach would be to pre-segment the video into different shots and compare both models on these shots. In fact this is the kind of approach adopted for many sports action recognition tasks using the HMM. However,

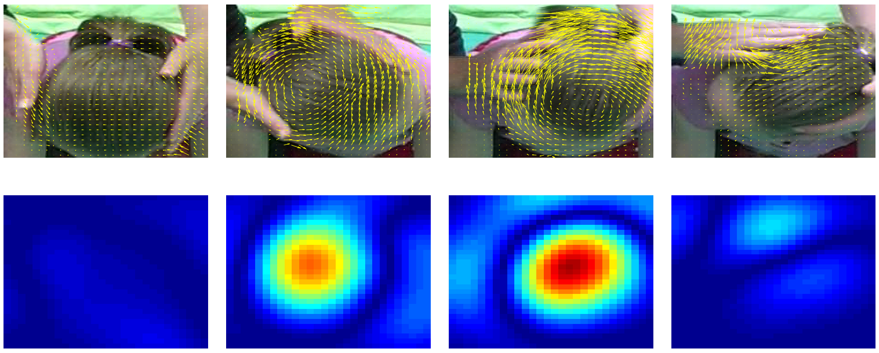


Fig. 10. The top four images show a selection of frames used to demonstrate a sequence of head rotation. The bottom four images show the sequence for the curl matrix. All of the above images have been zoomed in to improve clarity.

since both events are particularly hard to differentiate, this segmentation is not practical. A small variation in the use of the HMMs can however avoid pre-segmenting the video and allow to analyse the stream directly. Consider the layout of Figure 11. By stacking both HMMs in a single network of HMMs, it becomes possible to parse for $\mathcal{R}$ and $\overline{\mathcal{R}}$ simultaneously. Now for each frame of the video, the likelihood for the eight states of both HMMs is evaluated at the same time. The extra links between exit states S_8,S_4 and entry states S_1,S_5 are the glue which allows to switch between both models. They define how likely it is to switch from a rotation model to a non-rotation model, and vice-versa. Running Viterbi on this network of HMMs returns the MAP sequence of states, that, by looking at which HMM they belong to, can be simply translated in a sequence of $\mathcal{R}$ and $\overline{\mathcal{R}}$ events. This HMM framework thus does not simply classify previously parsed segments of video but jointly parses and classifies the events.

Twenty three exercise videos have been selected for evaluating this framework, totalling approximately 20 minutes of footage. All twenty three videos have rotational events manually noted for ground truths used in testing. Sixteen videos have been selected at random for training purposes and seven selected for testing. Both HMMs for $\mathcal{R}$ and $\overline{\mathcal{R}}$ are trained individually using the Baum-Welsh algorithm. The state transitions are reported on Figure 11 and the detail of the Gaussian distributions parameters are listed in [27]. The transitions between both models have been obtained by looking at the relative frequency of transitions between the models in the ground truths sequences. Note that these inter-model transitions can also be refined using an iterative Viterbi re-estimation scheme [30]. Note that different HMMs topologies have

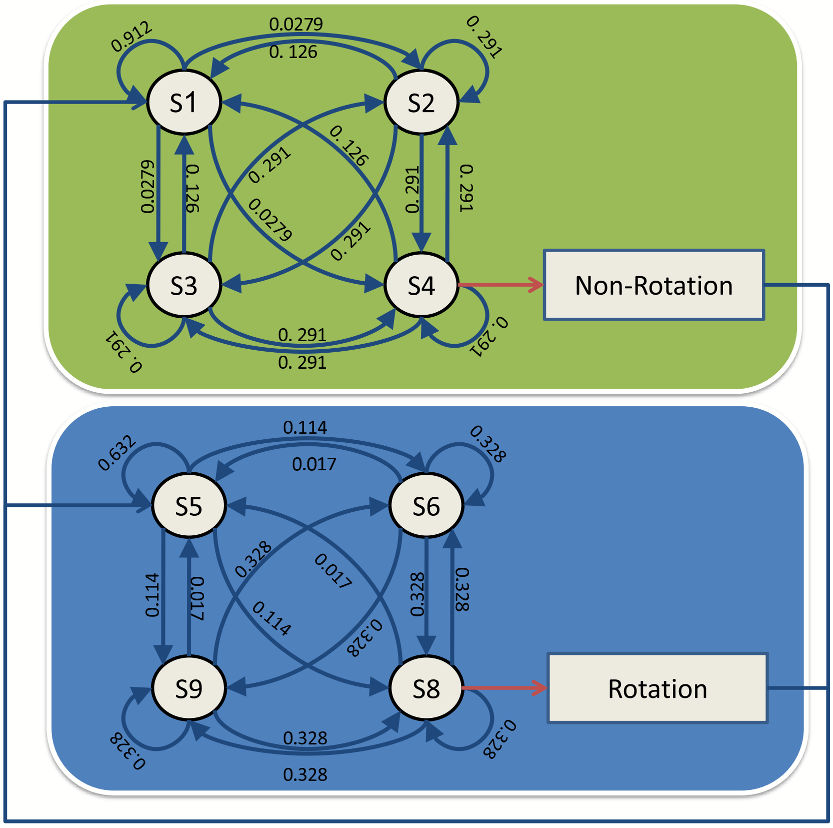


Fig. 11. Topology of the HMM network. On the top, the fully connected 4-state HMM for the non-rotational model, on the bottom the fully connected 4-state HMM for the rotational model. Both HMMs are linked to each other to allow a simultaneous segmentation of both models in the sequence.

been examined, and it seems that the fully connected 4-state model is optimal for this application.

The Viterbi algorithm has then been run using the two trained HMMs $\mathcal{R}$ and $\overline{\mathcal{R}}$ to recognise rotation events. The comparison between the estimates calculated by the network of HMMs and the manual segmentations is presented in Table 1. The table 1 reports the average Recall and Precision as well as the standard deviation of the Recall and Precision for all 23 video sequences, the 16 training sequences alone and the 7 testing sequences. A tolerance of 14 frames, roughly half a second, is allowed between the HMM estimates and manual segmentations. This is to allow for human error in noting rotation events, as a human observer can sometimes mis-classify pre-rotation head translation as rotation.

videos	Recall	Precision	Recall Standard Deviation	Precision Standard Deviation
All (23)	91.78	90.68	7.14	7.72
Training (16)	92.12	90.21	6.78	8.80
Test (7)	91	91.77	8.42	4.80

Table 1. Feature Evaluation: Hidden Markov Models Vs. Manual Markers. Recall, Precision.

4 Final comments on usability

To assess how usable in general this technology is, it is possible to seek evidence of exploitation of these ideas in everyday consumer equipment. No doubt a Tivo or Sky set top box would be the ideal place to exploit action metadata encoded into the transmitted sports bit stream, and behavioural psychologists attempting to use hundereds of hours of video would benefit from these ideas. However, right now, action spotting for the everyday consumer or scientific user is non existent. This would imply that the ideas are still new and not robust enough for operation in the marketplace. One of the main problems remains the generalisability of the algorithms. Direct parsing seems to work well, but in much of the published work, many more hours of testing seems to be necessary. In addition Direct parsing requires quite a deal of domain knowledge and the ideas seem to be very good for sports, but little else.

The future of action reconition in multimedia streams must therefore lie in the proper exploitation of dynamic inference engines like the HMM. In speech recognition, the use of statistical context-free grammar is widely spread [40]. We can imagine similar visual applications in which semantic parsing of videos without shot cut detection is possible. In a sense the community should aspire to the level of achivement of the speech recognition community. That community has benefitted greatly from the discovery of features (e.g. cepstral) which give good information for speech content. In a similar way the notion of visual words (e.g. as established by Zisserman et al [37]) could be exploited in an HMM for temporal parsing. This is certainly not a simple task but one step in that direction is more effort in unravelling the many subtleties of the HMM. Some discussion alonmg these lines is undertaken elsewhere in this book.

References

1. B. Alp, P. Haavisto, T. Jarske, K. Oistamo, and Y. Neuvo. Median based algorithms for image sequence processing. In *Visual Communications and Image Processing*, pages 122–134, 1990.
2. W. Arentz and B. Olstad. Classifying offensive sites based on image content. *Computer Vision and Image Understanding*, 94:295–310, 2004.
3. A. Bosson, G. Cawley, Y. Chan, and R. Harvey. Nonretrieval: blocking pornographic images. In *CIVR*, pages 50–60, 2002.
4. Yi Chan, Richard Harvey, and J. Andrew Bangham. Using colour features to block dubious images. In *European Signal Processing Conference (EUSIPCO)*, 2000.
5. P. Chang, M. Han, and Y. Gong. Extract highlights from baseball game video with hidden markov models. In *IEEE International Conference on Image Processing*, pages 609–612, September 2002.
6. J. Coopersmith. Pornography, videotape, and the internet. *IEEE Technology and Society Magazine*, pages 27–34, Spring 2000.
7. R. Coudray and B. Besserer. Global motion estimation for MPEG-encoded streams. In *International Conference on Image Processing*, pages 3411–3414, 2004.
8. R. Coudray and B. Besserer. Motion Based Segmentation using MPEG Streams and Watershed Method. In *International Symposium on Visual Computing*, pages 729–736, 2005.
9. R. Cutler and L.S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 22(8):781–796, August 2000.
10. R. Dahyot, A. C. Kokaram, N. Rea, and H. Denman. Joint audio visual retrieval for tennis broadcasts. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003.
11. R. Dahyot, N. Rea, A. Kokaram, and N. Kingsbury. Inlier modeling for multimedia data analysis. In *IEEE International Workshop on MultiMedia Signal Processing*, pages 482–485, Siena Italy, September 2004.
12. H. Denman, N. Rea, and A. Kokaram. Content-based analysis for video from snooker broadcasts. *Journal of Computer Vision and Image Understanding, Special Issue on Video Retrieval and Summarization*, 92:141–306, November/December 2003.
13. F. Dufaux and J. Konrad. Efficient, robust and fast global motion estimation for video coding. *IEEE Transactions on Image Processing*, 9:497–501, 2000.
14. A. Ekin and A. M. Tekalp. Automatic soccer video analysis and summarization. In *SPIE International Conference on Electronic Imaging: Storage and Retrieval for Media Databases*, pages 339–350, Jan 2003.
15. A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transaction on Image Processing*, 12(7):796–807, July 2003.
16. C. Fangxiang, W. Christmas, and J. Kittler. Periodic human motion description for sports video databases. In *International Conference on Pattern Recognition*, volume 3, pages 870 – 873, 2004.
17. Margaret M. Fleck, David A. Forsyth, and Chris Bregler. Finding Naked People. In *European Conference on Computer Vision (2)*, pages 593–602, 1996.

18. Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi. Automatic parsing of tv soccer programs. In *International Conference on Multimedia Computing and Systems*, volume 7, pages 167–174, May 1995.
19. Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.
20. E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot. Hmm based structuring of tennis videos using visual and audio cues. In *IEEE International Conference on Multimedia & Expo*, volume 3, pages 309–312, July 2003.
21. A. Kokaram and P. Delacourt. On the motion-based diagnosis of video from cricket broadcasts. In *Irish Signals and Systems Conference*, June 2002.
22. A. Kokaram, N. Rea, R. Dahyot, M. Tekalp, P. Bouthemy, P. Gros, and I. Sezan. Browsing sports video: trends in sports-related indexing and retrieval work. *IEEE Signal Processing Magazine*, 23, March 2006.
23. 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.
24. Anil Kokaram, Erika Doyle, Daire Lennon, Laurent Joyeux, , and Ray Fuller. Motion based parsing for video from observational psychology. In *Proc. SPIE, Multimedia Content Analysis, Management, and Retrieval*, volume 6073, 2006.
25. J. J. Lee, J. Kim, and J. H. Kim. Data-driven design of HMM topology for on-line handwriting recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1), 2001.
26. D. Lennon, N. Harte, and A. Kokaram. A HMM framework for motion based parsing for video from observational psychology. In *Irish Machine Vision and Image Processing Conference*, pages 110–117, DCU, Dublin ,Ireland, August 2006.
27. Daire Lennon. Motion based parsing. Master's thesis, Trinity College Dublin, 2007.
28. R. Leonardi, P. Migliorati, and M. Prandini. Semantic indexing of soccer audio-visual sequences: A multimodal approach based on controlled markov chains. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5), May 2004.
29. Che-Bin Liu and Narendra Ahuja. Motion based retrieval of dynamic objects in videos. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 288–291, 2004.
30. J. Odell, D. Ollason, P. Woodland, S. Young, and J. Jansen. *The HTK Book for HTK V2.0.* Cambridge University Press, Cambridge, UK, 1995.
31. J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4), December 1995.
32. F. Pitié, S-A. Berrani, R. Dahyot, and A. Kokaram. Off-line multiple object tracking using candidate selection and the viterbi algorithm. In *IEEE International Conference on Image Processing (ICIP'05)*, Genoa, Italy, 2005.
33. N. Rea, R. Dahyot, and A. Kokaram. Semantic event detection in sports through motion understanding. In *3rd International Conference on Image and Video Retrieval (CIVR 04)*, Dublin, Ireland, July 2004.
34. N. Rea, R. Dahyot, and A. Kokaram. Classification and representation of semantic content in broadcast tennis videos. In *IEEE International Conference on Image Processing (ICIP'05)*, Genoa, Italy, 2005.

35. N. Rea, C. Lambe, G. Lacey, and R. Dahyot. Multimodal periodicity analysis for illicit content detection in videos. In *IET 3rd European Conference on Visual Media Production (CVMP)*, pages 106–114, London, UK, November 2006.
36. W. A. Sethares and T. W. Staley. Periodicity transforms. *IEEE transactions on Signal Processing*, 47(11), November 1999.
37. J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.
38. M. Vlachos, P. Yu, and V. Castelli. On periodicity detection and structural periodic similarity. In *SIAM International Conference on Data Mining*, 2005.
39. J. Ze Wang, J. Li, G. Wiederhold, and O. Firschein. System for screening objectionable images using Daubechies' wavelets and color histograms. In *International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services*, pages 20–30, 1997.
40. S.J. Young, N.H. Russell, and J.H.S Thornton. Token passing: A simple conceptual model for connected speech recognition systems. Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Dept, 1989.