Bayesian Fusion of Hidden Markov Models for Understanding Bimanual Movements

Atid Shamaie Alistair Sutherland
Centre for Digital Video Processing, School of Computing,
Dublin City University, Dublin 9, Ireland.
{ashamaie, alistair}@computing.dcu.ie

Abstract

Understanding hand and body gestures is a part of a wide spectrum of current research in computer vision and Human-Computer Interaction. A part of this can be the recognition of movements in which the two hands move simultaneously to do something or imply a meaning. We present a Bayesian network for fusing Hidden Markov Models in order to recognise a bimanual movement. A bimanual movement is tracked and segmented by a tracking algorithm. Hidden Markov Models are assigned to the segments in order to learn and recognize the partial movement within each segment. A Bayesian network fuses the HMMs in order to perceive the movement of the two hands as a single entity.

1. Introduction

Bimanual movements form a large volume of daily human movements. Clapping, opening a bottle, typing on a keyboard, eating with knife and fork, drumming, showing the size of something, etc. are some typical bimanual movements.

The problem of recognising hand gestures has been widely addressed in the literature [1]. Gesture recognition by Hidden Markov Models [2], parametric Hidden Markov Models for recognition of hand gestures [3], spatio-temporal hand gesture recognition using neural networks [4], HMM-based threshold models for gesture recognition [5], and many other techniques have been used to deal with the problem of hand gesture recognition.

Coupled Hidden Markov Models have been proposed in the literature to deal with the simultaneous movements of the two hands [6]. However, a major weakness of Coupled HMM is that this model is unable to deal with occlusion properly. In this model, the hands must be separately recognisable throughout the whole image sequence. In fact they use colour gloves with different colours to extract the hands in the images. In a more realistic situation where we do not wear gloves, we cannot ignore occlusion parts of the movements where one hand covers the other partially or completely. In addition, in some movements one hand may be occluded by another object (e.g. the body) or leave the scene. A general solution must consider all types of occlusion.

In this paper we do not use colour gloves. Therefore, during an occlusion period both hands are extracted as a single blob in the images.

A prerequisite in bimanual movement understanding is hand tracking. Since one hand may cover the other occasionally during the movements a tracking algorithm is needed to track the hands and reacquire them at the end of each occlusion period. A Bayesian network-based technique for tracking interacting hands [7] and a dynamic model for hand tracking [8] have been proposed in the literature.

In this paper, we introduce a novel Bayesian network for recognising bimanual movements that is capable of dealing with some fundamental issues such as occlusion and disappearance of one hand. First, we review a tracking algorithm. In Section 3, we do gesture segmentation in order to separate the occlusion parts from the non-occlusion parts of a bimanual movement. Section 4 is dedicated to the Bayesian network and fusion of Hidden Markov Models for understanding a movement. Experimental results are presented at the end of paper.

2. Hand tracking and bimanual coordination

We use a hand tracking algorithm based on a neuroscience phenomenon called *bimanual coordination* and Kalman filtering [8]. In this algorithm, the hands are tracked in a sequence of images and occlusion is detected by a dynamic model and Kalman filtering. Also the tracker is able to track the hands when one of them leaves the scene and return. It distinguishes between the occlusion and the presence of one hand at different stages throughout the processing period.

Bimanual coordination implies that the hands are highly synchronised both spatially and temporally. Therefore, the behaviour of the hands can be understood by considering this restriction. During an occlusion

Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04)
0-7695-2122-3/04 $ 20.00 © 2004 IEEE

period the tracking algorithm detects hand pauses by monitoring the hands velocities. Due to bimanual coordination the pauses in the movements of the two hands happen simultaneously. The tracking algorithm tracks the hands during an occlusion period by distinguishing the case where the hands pause or collide and return to their previous position (e.g. in clapping) from the case where the hands pass each other without any pause or collision. Due to the fact that the hands can have many different types of movements a set of models are defined to model almost every movement of the hands in bimanual movements [8]. The hands are said to be positively synchronised when they move in the same direction. When the hands move in different directions they are said to be negatively synchronised. Using synchronisation and hand pause detection models, the tracking algorithm reacquires the hands at the end of each occlusion period. The tracking algorithm is independent of the hand shapes [8]. Therefore, it allows the shapes to change during the movements and occlusion periods.

3. Movement segmentation

A bimanual consists of the "occlusion parts", in which one hand occludes the other, and the "non-occlusion" parts, where the hands are recognisable separately. Since a bimanual movement can be essentially a periodic movement like clapping, we separate different parts, which we call "segments". Four segments are obtained as follows,

A. The beginning segment, from the beginning of a movement to the first occlusion part

B. The occlusion segments, where one hand is occluded

C. The middle segments, a part of a movement between two consecutive occlusion segments

D. The end segment, from the last occlusion segment to the end of movement

An example of a segmented bimanual movement is shown in Figure 1. Although, we have assumed that the movements start and end in non-occlusion segments, extending the recognition algorithm to the other cases is straightforward and makes no difference in the essence of the algorithm. For example, in a movement where the hands stop moving in an occlusion segment there will be no end segment. Also for the movements in which no occlusion segment is observed the process is the same with only one segment for the whole movement. We will see in the next section that the algorithm recognises the segments and consequently the whole movement so that they can end in an occlusion segment or even in the beginning segment (in the case of only one segment).

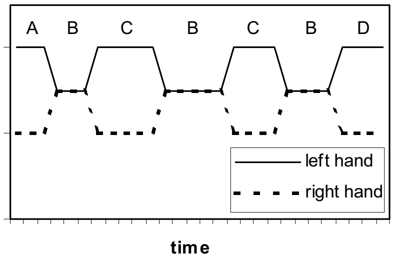


Figure 1. A segmented bimanual movement. This figure shows the movement of the hands in non-occlusion and occlusion segments. Where the two lines overlap one hand is occluded partially or completely. In the non-occlusion segments the lines have no overlap.

4. Bayesian fusion of Hidden Markov Models

A Hidden Markov Model (HMM) is assigned to each hand in each non-occlusion segment. A separate HMM is assigned to the occlusion segments. Therefore, in a typical bimanual movement, seven separate HMMs are assigned to the movement, two for the beginning segments for the two hands, one for the occlusion segments, two for the middle segments, and two for the end segments. The HMMs are trained to recognise the partial movement of the hands in each segment. For example, in a middle segment of an unknown movement, the trained HMMs of the middle segments of each hand in the vocabulary are employed to calculate the likelihoods in the given middle segment. The likelihoods are then forwarded to a Bayesian network in which they are fused to recognise the partial movement of the hands in the given segment.

4.1. A Bayesian network for recognition

In a periodic bimanual movement like clapping there can be several occlusion and middle segments. For example, in Figure 1, there are 3 occlusion and 2 middle segments. Thus, a data fusion structure must be able to deal with multiple occlusion and middle segments as well as the beginning and ending segments in order to recognise the whole movement. A well-known method of data fusion is the Bayesian network. Bayesian networks are suitable for tasks of analysis (e.g. medical diagnosis) to piece together a model of physical reality [9]. The other methods of data fusion such as Dempster-Shafer are more suitable for a synthesis type of task where the constraints are imposed externally[1].

We introduce a Bayesian network in which a bimanual movement is divided into the movements of the two hands. The movement of each hand is also divided into occlusion and non-occlusion segments (see Figure 2). In this network the LH and RH nodes stand for the

[1] Many data fusion methods such as Depster-Shafer, Fuzzy Logic, and Neural Networks have been addressed in the literature that requires a separate research for this particular problem.

movement of the left hand and the right hand respectively. The BG node stands for the bimanual gesture, the NS nodes are the evidence nodes for the non-occlusion segments and the OS is the evidence node for the occlusion segments. Since an occlusion segment is a common part for both hands a single shared node is considered for this segment. Also, due to the fact that the beginning, middle, and end segments have no time overlap, and assuming that the segments are of equal weight, one NS node per hand is all that is required.

At the lowest level of the network are the HMMs of the partial movements of the hands in the associated segments. The HMMs of the partial movements in each segment are fused so that the whole bimanual movement is recognised at the root node. Each node in the causal tree represents a multi-valued variable. The NS and OS evidence nodes provide evidence to both LH and RH nodes. These evidence nodes are fed by the HMMs of different segments separately. The set of trained HMMs of every segment is employed to calculate the likelihoods of a given segment of a gesture [10]. These values represent the likelihoods that the given gesture is each of the gestures in the vocabulary in the period of the given segment. Therefore, every evidence node represents a normalised vector of likelihoods. The vectors are, then, used as evidence messages in Pearl's belief propagation algorithm [9] for updating the local belief of the nodes in the network.

This level of separation between the individual movements of the hands enables us to keep recognising the movements in which one hand leaves the observable scene occasionally. In this case, the sub-tree of the visible hand including the corresponding NS node and the HMMs is the only source of information feeding to the network.

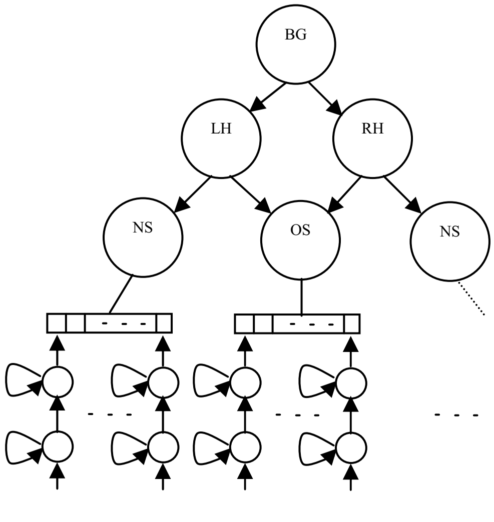


Figure 2. The structure of the proposed Bayesian network. The HMMs at the lower part calculate the vectors of likelihoods which are forwarded to the causal tree at the upper part.

The structure of the network seems to contain a loop. The nodes BG, LH, OS, and RH form a loop. However, in our network the OS is an evidence node which does not receive messages and it always transmits the same vector. Therefore, messages cannot circulate and, in fact, the loop is cut by the evidence node.

4.2. The Hidden Markov Models

For each hand in a non-occlusion segment of a movement a 2-state left-to-right HMM (shown on the bottom of the network of Figure 2) is trained[2] by using a set of training data, consisting of videos of bimanual movements.

Since we do not separate the hands in an occlusion segment, the image sequence of the segment is treated as a single segment and a HMM is assigned to it. While in the other segments each hand is extracted separately and a HMM is assigned and trained for each hand.

The normalised vector of likelihoods calculated by the HMMs in each segment is forwarded to the evidence nodes as the vector of probabilities that the given movement is each of the movements in the vocabulary in the given segment. Therefore, each vector has the same number of elements as the number of movements in the vocabulary. For example, if the network is trained for g bimanual movements, each evidence vector has g elements. Also, each node of the causal tree represents a vector of length g.

The messages propagate through the network and the movement is recognised as the movement with the highest probability at the root node.

The algorithm for recognising bimanual movements is summarised as follows:

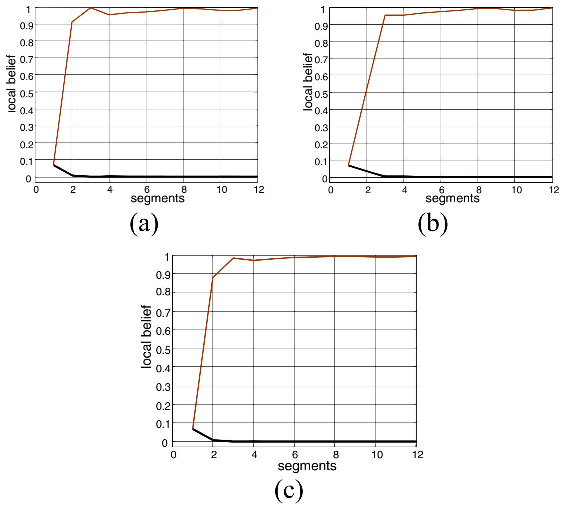
```
1.  A bimanual movement is segmented by the tracking
    algorithm

2.  The beginning segment
    2.1. Using the HMMs of the beginning segment of
         each hand the vector of likelihoods is calculated
         and normalised
    2.2. The vectors of likelihoods are passed into the
         corresponding NS nodes while the vector of
         occlusion node is set to a vector of all 1s.
    2.3. The nodes' beliefs are updated

3.  An occlusion segment
```

[2] The features space of the HMMs is constructed with the seven principal components of the hand shapes and the movement vector. An eigenspace is made for each of the two hands and for the hand-hand overlap in the occlusion segments. The principal components are the projection of images of the hands into the corresponding eigenspaces. The movement vector represents the direction of the movement of a hand in two consecutive images.

```
    3.1. The vector of likelihoods is calculated and
         normalised by using the occlusion segments
         HMMs
    3.2. The vector is passed into the OS node
    3.3. The nodes' beliefs are updated

4. A middle segment
    4.1. The vectors of likelihoods are calculated and
         normalised by using the corresponding HMMs
         of the middle segments
    4.2. The vectors of likelihoods are passed to the
         corresponding NS nodes
    4.3. The nodes' belief are updated

5. While there are more occlusion and middle segments
   the parts 3 and 4 of the algorithm are repeated

6. The end segment
    6.1. The vectors of likelihoods are calculated and
         normalised by using the HMMs of the end
         segment
    6.2. The vectors are passed to the corresponding NS
         nodes
    6.3. The nodes' beliefs are updated

7. The movement with the highest probability in the
local belief of the root node is the best match
```

4.3. Experimental results

15 bimanual movements were created as if the hands were doing regular daily movements like clapping, knotting a string, turning over the leaves of a book, and some movements from British Sign Language. In two different conditions, the camera was positioned at two distances and heights from the subject with less than 50cm difference in two different rooms to record a set of videos of the movements. 33% of the videos were recorded at the shorter distance in the room number 1 and the rest at the longer distance in the room number 2. For every movement 10 videos were captured 5 of each were treated as the training set and the rest as the test set. The main sources of variations recorded in the videos are different camera positions, different initial hand positions, rotations and shapes, variations in performing the examples of each movement, and different places of video recording.



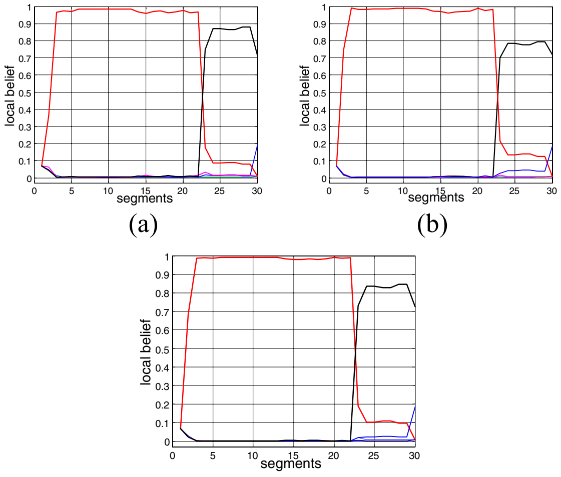
(a) (b)

(c)

Figure 3. The belief changes of the (a) LH, (b) RH and (c) root node, BG, for a clapping movement from the test set.

Here we present an example from the test set. Figure 3 shows the belief change of the nodes LH, RH, and BG for a clapping from the test set. The belief of the nodes starts from the initial equilibrium. These initial points are presented as the initial segment in the graphs of Figure 3. By processing the beginning segment of the movement the beliefs of the nodes are updated and presented as the second segment in the graphs. The graphs show that the movement has been recognised rapidly in the beginning segments and this result has been preserved throughout the rest of movement.

In order to evaluate the algorithm and the network, they were employed to recognise all the movements in the test set. 74 out of 75 movements in the test set were recognised correctly. The graphs of the belief change of the only movement that was not recognised correctly are sketched in Figure 4. At the beginning the movement was correctly recognized in all the three nodes of LH, RH, and BG. From one point onward the beliefs are changed so that the movement was recognised differently. Our investigation shows that from this point the HMMs have resulted in different likelihoods and this result has been preserved throughout the rest of movement. Although, in the mis-recognised part of the movement the beliefs are not as confident as the correctly recognised part, recognition is performed based on the highest probability. Therefore, it is concluded that the recognised movement is the movement with highest probability in the local belief of the root node.



(a)

(b)

(c)

Figure 4. The belief changes of the movement recognised wrongly at the (a) LH, (b) RH, and (c) BG nodes.

4.4. Recognition of concatenated periodic bimanual movements

Many bimanual movements can be periodic in essence. Clapping and drumming are some examples. In environments where bimanual movements are used as a communication method, e.g. Virtual Reality, recognising concatenated periodic movements is crucial.

We employ the proposed Bayesian network in order to recognise this type of movement. The important points in such a process are correct recognition of the movements over the whole periods and exact detection of movement changes when different movements are concatenated. We use the trained models of the last section.

Three bimanual movements were performed consecutively, each of which was repeated dozens of times. From the 15 movements, first movement number 3 was repeated 5 times. It was followed by movement number 2 repeated 30 times and followed by movement number 5 repeated 41 times. Therefore, the first movement is divided into 11 segments, the second one into 61 segments, and the last one into 83 segments. Given the fact that the first segment in the graph of local beliefs represents the belief of initialisation, the first movement transition should appear in the 13th segment and the second transition in the 74th segment. The local belief of the root node is plotted in Figure 5. The movements are correctly recognised most of the time. Also, the transitions are detected properly. However, it can be seen, particularly in the graph of the second movement, that the belief is not very stable and it varies such that at some points it falls below the graph of other movements. This happens when the partial movements of one or two hands are recognised wrongly.

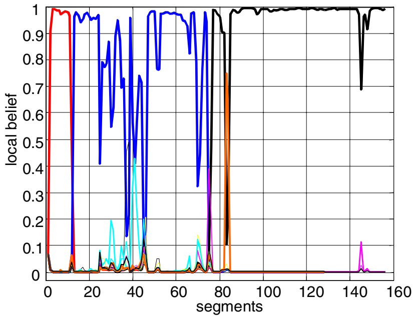
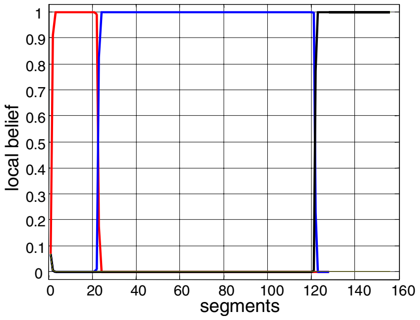


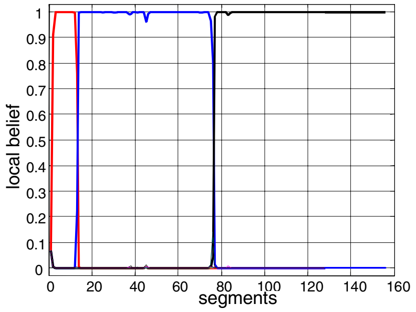
Figure 5. The belief change of the root node for three concatenated movements. Each colour line represents the belief of the root node for a particular gesture.

Although, the confusions can be treated as temporary spikes, we may come to a conclusion that the movement has changed at some points. In fact the beliefs of other movements are larger than the second movement at those points that support the transition hypothesis.

In order to avoid the confusing spikes, we replace the prior probability of the root node at each belief updating stage with its current local belief subject to a limitation. The applied limitation does not allow the prior probabilities of the root node to become smaller than a threshold. If we do not apply this restriction, repetition of the movements can cause very small probabilities resulting in extreme delays in detecting transition points. Figure 6a shows the result of the algorithm without the restriction. Obviously, the beliefs are stable while the transition points have been detected 9 and 47 segments later than the actual transition points respectively. However, by applying a threshold of 10^{-3} the local belief of the root node is stabilised and the transition delays are reduced to 1 and 2 segments respectively (see Figure 6b).



(a)



(b)

Figure 6. The local belief of the root node is stabilised. (a) the network without restriction results in extreme transition delays, (b) the network with restriction results in short transition delays.

4.5. Comparisons

The presented Bayesian network and algorithm for learning and recognition of Bimanual movements has several advantages over the Coupled Hidden Markov Models. It can deal with the occlusion problem in a realistic situation where no colour glove is used. Given that in the experiments reported in [6] the hand shapes are ignored, with a vocabulary of 3 T'ai Chi bimanual gestures Coupled HMM has resulted in 94.2% recognition rate using 1/3 of the test set as the training set. The performance of 74/75 (~98%) recognition rate presented in this paper includes a vocabulary of 15 bimanual movements without using the test set for the training purpose.

5. Conclusion

A new technique for recognition of bimanual movements was presented. We used a tracking algorithm to track the hands and do movement segmentation all over a bimanual movement including occlusion. A Bayesian network was proposed for recognising bimanual movements. Our experimental results demonstrated significant performance of the network given that it can deal with occlusion. The proposed network was also employed to recognise concatenated periodic bimanual movements. We made a slight change in the belief propagation algorithm to stabilise the belief of the network. The advantages of the proposed Bayesian network and bimanual movement recognition algorithm were discussed.

6. References

[1] R. Cipolla and A. Pentland, *Computer Vision for Human-Machine Interaction*, Cambridge University Press, 1998.

[2] T. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models", *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, 1995.

[3] A.D. Wilson and A. Bobick, "Parametric Hidden Markov Models for Gesture Recognition" *IEEE Trans. Pattern Analysis Machine Intelligence,* Vol. 21, No. 9, 1999, pp. 884-900.

[4] M. Su, H. Huang, C. Lin, C. Huang, and C. Lin, "Application of Neural Networks in Spatio-Temporal Hand Gesture Recognition", Proc. IEEE World Congress on Computational Intelligence, U.S.A., 1998.

[5] H. Lee and J. Kim, "An HMM-Based Threshold Model Approach for Gesture Recognition", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 21, No. 10, October 1999.

[6] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition", Proc. Conf. Computer Vision and Pattern Recognition, Peurto Rico, 1997.

[7] S. Gong, J. Ng, and, J. Sherrah, "On the Semantics of Visual Behaviour, Structured Events and Trajectories of Human Action" *Image and Vision Computing,* Vol. 20, 2002, pp. 873-888.

[8] A. Shamaie and A. Sutherland, "A Dynamic Model for Real-Time Tracking of Hands in Bimanual Movements" *5th International Gesture Workshop*, Genova, Italy, April 2003.

[9] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*, California, Morgan Kaufmann Publishers, 1998.

[10] Z. Ghahramani, "An Introduction toHidden Markov Models and Bayesian Networks", in *Hidden Markov Models: Applications in Computer Vision*, World Scientific, Singapore 2001.