Multiple Frame Motion Inference Using Belief Propagation

Jiang Gao
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jgao@cs.cmu.edu

Jianbo Shi
Department of Computer and Information Science
University of Pennsylvania
Philadelphia,PA 19104
jshi@cis.upenn.edu

Abstract

We present an algorithm for automatic inference of human upper body motion. A graph model is proposed for inferring human motion, and motion inference is posed as a mapping problem between state nodes in the graph model and features in image patches. Belief propagation is utilized for Bayesian inference in this graph. A multiple-frame inference model/algorithm is proposed to combine both structural and temporal constraints in human motion. We also present a method for capturing constraints of human body configuration under different view angles. The algorithm is applied in a prototype system that can automatically label upper body motion from videos, without manual initialization of body parts.

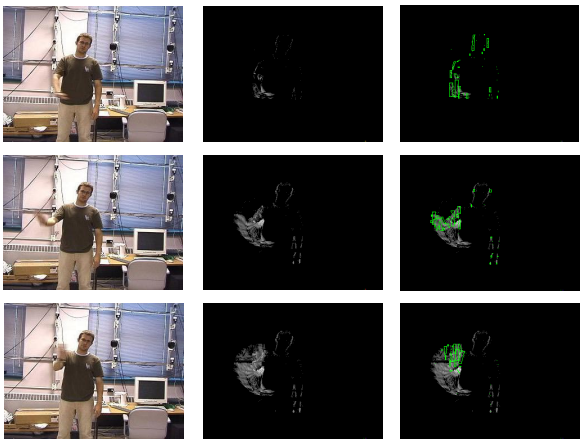
1. Introduction

Human motion detection and tracking has many applications. For example, motion perception in a human-machine interface could enable people to communicate with computers using body language or gestures. Another application is human activity analysis, in which human motion and gestures are detected and recognized from surveillance cameras.

Many research activities have been directed toward tracking and recognition of human motion and gestures. In this paper, we describe our approach for automatic inference of human upper body motion from motion energy images and color features.

1.1. Previous works

While many works have been done on human motion tracking (Bregler 1998, Ju 1996), most of the algorithms need manual initialization of body parts for tracking. For algorithms with self-initialization ability, only some of them estimate details of body parts.



(a) (b) (c)

Figure 1. Motion energy images of a gesture with motion history accumulated. (a) Input frames. (b) Accumulations of motion energy images starting from the 1st frame. (c) Same as (b), with motion energy pixels of the current frame marked.

Generally, automatic labeling of body parts is based on selected image features and techniques. Background subtraction is an effective technique for human detection and tracking (Haritaoglu 1998, Felzenszwalb 2001), and are widely used in video surveillance applications with static cameras.

The Multi-view approach (Gavrila 1996) makes use of 3D information and can resolve some of the ambiguities in complex situations, such as occlusions. The application of background subtraction and multi-view algorithms may not always be possible in some applications, such as instant human activity analysis in single camera videos.

Body structure approach is proposed recently based on component models of human body (Ioffe 2001, Felzenszwalb 2001). To label body parts, the algorithms search the space of possible human configurations, and find the best match with image

observations. In Mori (2002), shape context matching is used to match contours of body parts.

In this paper, we propose a new framework that can infer human upper body motion and label body parts without manual initialization. We pose body parts labeling as a Bayesian inference problem in a Markov network (Jordan 1998, Yedidia 2001). Our motivation and method are similar with Freeman (2000), though with different applications. Our model is proposed to capture constraints of human body configuration under different view angles. A multiple-frame Markov network model is further proposed for combining both temporal and structural constraints in the Markov network. We use belief propagation, which performs spatial and temporal inference simultaneously, to infer body motion in the Markov network. We are using this approach to design an intelligent human machine interface, where we can assume limited view angles, single person, and still background.

1.2. Motion energy images

Motion-energy image (MEI) is a motion feature for representing moving regions (Bobick 2001). Let $D(x, y, t)$ be a binary image sequence indicating regions of motion. $D(x, y, t)$ can be obtained by image differencing followed by a thresholding. In this paper we assume

$$D(x, y, t) = 1 \quad (1)$$

represents the pixel at (x, y) in frame t is in motion, then the binary motion-energy image $E_\tau(x, y, t)$ is defined as

$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t-i) . \quad (2)$$

τ is the temporal duration for computing the MEI.

Fig. 1 shows an accumulation of MEI of a gesturing activity in several frames. The outlined pixels in Fig.1 (c) are the current slice of motion energy pixels.

In this paper, we use motion-energy image as the motion feature to infer 3D position of body joints. To do this, we proposed a Markov network model to embed constraints of human body structure. We also propose a multiple frame Markov network model to take advantage of temporal consistency in human motion.

We divide the task into 2 phases. First, inferring 2D positions of body joints in images. Second, recover 3D positions of body joints. In the following, we will first present our inference model. The method for combining motion and color features into the inference models is given in section 4.

The organization of this paper is as follows. In section 2, we describe a Markov network model for human upper body motion inference in a single frame. In section 3, we propose a multi-frame model for Bayesian inference. Section 4 describes using motion and color features to constrain the inference space, and gives experimental results. Finally, section 5 concludes the paper.

2. Modeling single frame probability

Our goal is to infer positions of body joints based on motion energy images. In Bayesian framework, given image features *x*, body configuration *h* can be estimated as:

$$h = \arg\max_h P(h \mid x) , \quad (3)$$

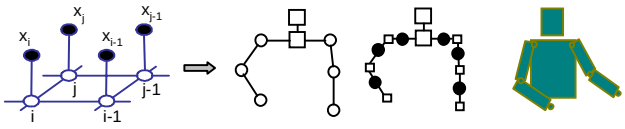
$$P(h \mid x) = cP(x \mid h)P(h) . \quad (4)$$

Here body configuration *h* consists of positions of body joints, denoted in this paper as $(s_1, s_2, \cdots s_N)$.

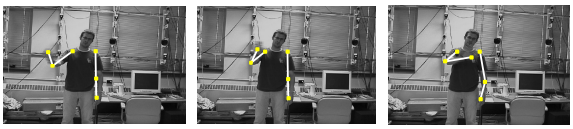
2.1. The Markov network model

We propose the Markov network model as shown in Fig. 2(a) to solve Eqs. (3) and (4). In this model state nodes (the empty circles) represent 2D positions of body joints. In the Markov network model each state node is connected with a measurement node (the filled-in circles), as well as to its neighbors.

We denote a state at node *i* as s_i, and observation at the corresponding measurement node as x_i. As shown in Fig.2 (b), x_i corresponds to the body parts between joints. We define the image patches (observations) corresponding to wrist joint, elbow, and shoulder joint, as lower arm, upper arm, and shoulder girdle, respectively. The image patches are defined based on a cardboard person model (Fig. 2(c)).



(a) (b) (c)



(d)

Figure 2. (a)-(b) A Markov network model for upper body motion inference. The empty circles are state nodes, and filled-in circles are measurement nodes. (c) Cardboard person model for evidence computation. (d) An upper body motion inference result.

In this model, $P_{ij}(s_i, s_j)$ is the probability that two body joint positions appear together. $P(x_i \mid s_i)$ is computed by counting the number of motion energy pixels (x, y) in each image patches, i.e., $D(x, y, t)$ =1 (Eq.(1)). The definition is equivalent to say that the more motion energy pixels in the image patches, the more likely body joint positions defining the body part are right.

Clearly, $P(x_i \mid s_i)$ computed above does not directly correspond to a probability function. We need to convert this "energy" measurement to a probabilistic measurement. This is done by a transform:

$$P(x_i \mid s_i) = C / \{1 + \exp(-E)\}, \quad (5)$$

where E the number of motion energy pixels in an image patch, and C is a normalization coefficient.

The Markov network model essentially decomposes Eqs. (3)-(4) by:

$$h = (s_1, \cdots, s_N) = \arg\max_h P(h \mid x), \quad (6)$$

$$P(x \mid h = (s_1, \cdots, s_N)) = \prod_{i=1}^{N} P(x_i \mid s_i), \quad (7)$$

$$P(s_1, \cdots, s_N) = \frac{\prod_{s_i, s_j} P_{ij}(s_i, s_j)}{\prod_{s_i} P_i(s_i)^{\deg(s_i)-1}}, \quad (8)$$

where degree of s_i, $\deg(s_i)$ is the number of nodes connecting with s_i.

Eqs. (6)-(8) are solved by inference in the Markov network model. In the following, we first propose a learning algorithm to estimate parameters of the Markov network model, then describe an inference algorithm based on belief propagation.

2.2. Learning the Markov network model

The Markov network defined above decomposes the Bayesian inference problem Eqs. (3)-(4) into local states and their corresponding measurements or evidences. However, we need to estimate $P_{ij}(s_i, s_j)$ before we can do inference in this network. $P_{ij}(s_i, s_j)$ represents a priori probability for body joint positions, or configuration constraints of human body.

In this paper, we assume the face position can be estimated beforehand using algorithms such as face detection. The approximate position of shoulder girdle is then estimated based on position of face and pose assumptions. Since we are only interested in human upper body configuration, we need estimate the $P_{ij}(s_i, s_j)$'s between wrist, elbow, and shoulder joints.

We model joint probabilities of 2D projection of body joints, rather than directly model the constraints in 3D. The advantage of this approach is that it simplifies the modeling process and avoids recovering 3D pose at the beginning. The drawback is the modeling is view-specific. In our system, we train a separate set of $P_{ij}(s_i, s_j)$ for each different view angles. We consider only 3 view angles at this stage, namely, frontal, turning to the left, and turning to the right. For our application of human computer interface, the 3 view-angle assumption is enough.

We use a supervised learning method for estimating $P_{ij}(s_i, s_j)$ between 2D positions of joints i and j. We uniformly sample the 2D image space, as shown in Fig. 3, and only estimate joint probabilities at the sampling positions (intersections of the grid in Fig. 3). All the other positions are tied to the nearest sampling coordinates.

Before the learning process, each sampling position (s_i, s_j) are given the same probability:

$$P_{ij}(s_i, s_j) = 1.0 / (N_s * N_s), \quad (9)$$

where N_s is the number of sampling positions. Then we run our body part labeling system through video sequences. For each estimated pair of body joint positions (s_i, s_j) that is not a valid human body configuration, we reduce its probability by

$$P_{ij}(s_i, s_j) = P_{ij}(s_i, s_j) / T, \quad (10)$$

where T is a constant and $T > 1$. Then $P_{ij}(s_i, s_j)$ is renormalized by:

$$P_{ij}(s_i, s_j) = \frac{P_{ij}(s_i, s_j)}{\sum_{s_i, s_j} P_{ij}(s_i, s_j)}. \quad (11)$$

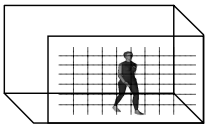


Figure 3. Learning joint probabilities at sampling points of a 2D image plane.

3. Modeling multiple-frame probabilities

In everyday life, people use temporal constraints of human motion trajectories to help tracking of body parts. While people may not be able to find body configuration at an instance with enough confidence, they can definitely track human body parts after a long sequence of human performance.

In this section, we extend the single frame Markov network model of section 2 into a Markov network for multiple frames of human motion.

3.1. Temporal Constraints

The temporal constraints are added to the Markov network model by define the joint probability of state nodes corresponding to the same body joint in consecutive frames, as shown in Fig. 4. Assuming s_i^t is a state of node i in frame t, and s_i^{t+1} is a state of node i in frame t+1, their joint probability is defined as follows:

$$P_i^{t,t+1}(s_i^t, s_i^{t+1}) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{|s_i^t - s_i^{t+1}|^2}{\sigma^2}, \quad (12)$$

which is a Gaussian distribution of 2D distance between s_i^t and s_i^{t+1}. The covariance parameter σ is determined empirically.

The joint probability (12) only imposes the smoothness of transition between s_i^t and s_i^{t+1} without any specific model. This gives the system ability to

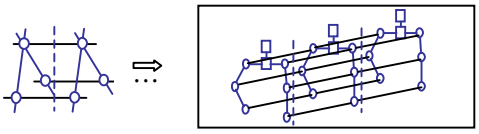


Figure 4. Adding temporal constraints to the model. By connecting the state nodes, states of the same body joint in consecutive frames are given a joint probability.

deal with a wide range of human motions. With temporal constraints, the Bayesian inference algorithm is more robust, and can even recover from labeling errors in a single frame.

3.2. Belief propagation

Belief propagation (BP) is an iterative algorithm to infer the hidden states (or solving Eqs. (6)-(8)) in a Markov network based on message passing. A basic iteration is as follows:

$$m_{ij}(s_j) = \alpha \sum_{x_i} P_{ij}(s_i, s_j) P_i(x_i \mid s_i) \prod_{k \in N(i) \setminus j} m_{ki}(s_i), \quad (13)$$

$$b_i(s_i) = \alpha P_i(x_i \mid s_i) \prod_{k \in N(i)} m_{ki}(s_i), \quad (14)$$

where m_{ij} is the message that node i sends to node j, and b_i is the belief, i.e., marginal posterior probability, at node i. b_i is obtained by multiplying all incoming messages to the node by the local evidence (likelihood). α is a normalization constant. $N(i) \setminus j$ means all nodes neighboring node i except j. All messages $m_{ij}(s_i)$ are initialized to 1 before the iterations begin.

Though belief propagation algorithm is exact (i.e., guaranteed convergence to the optimal solution) only in networks without loops, recent study shows that it can also converge in many loopy networks. Our multi-frame Markov network contains loops. It is therefore interesting to see if the BP algorithm can converge to optimal solution in this network.

As described in section 2.2, we defined three different view angles (poses). We compute the beliefs of possible body configurations for each pose. Body pose and configuration are determined simultaneously by selecting the one which has the highest belief given the observations.

3.3. 3D body configuration recovery

Recovering 3D configurations based on 2D projection of body joint positions is based on the algorithm of Taylor (2000). Assume (u_1, v_1) and (u_2, v_2) are projections of the 3D points (X_1, Y_1, Z_1) and (X_2, Y_2, Z_2) on the image plane, under orthographic projection we have

$$(u_1 - u_2) = s(X_1 - X_2), \quad (15)$$

$$(v_1 - v_2) = s(Y_1 - Y_2). \quad (16)$$

and it can be derived that

$$Z_1 - Z_2 = \sqrt{l^2 - ((u_1 - u_2)^2 + (v_1 - v_2)^2)/s^2}, \quad (17)$$

where s is a constant, and l is the length between (X_1, Y_1, Z_1) and (X_2, Y_2, Z_2). By assumimg a reference depth Z_0, and using the depth difference computed in (17), we can recover 3D positions of all the body joints. For details, refer to Taylor (2000).

4. Experimental results

4.1. System architecture

We developed a prototype system that can automatically detect and label human upper body motion in a natural environment. The algorithm is shown in Fig. 5.

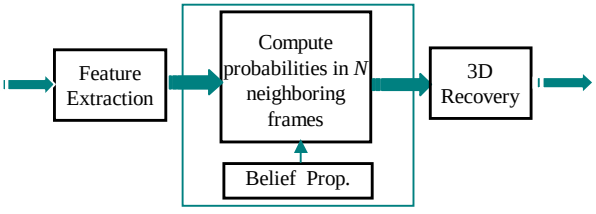


Figure 5. Upper body motion inference system.

The algorithm proceeds as follows: First, face detection is conducted; color and motion features are extracted (Section 4.1.1), and candidate positions for each body joints are detected based on the image features (Section 4.1.2). Then, Bayesian inference is conducted in the multi-frame Markov network model. We use belief propagation to find the best body configuration. At this stage, the estimated body configuration is 2D positions of body joints on images. We then apply 3D recovery algorithm to recover 3D coordinates of body joints (Section 3.3).

4.1.1. Color and motion features. We use 2 different features in our system. For motion feature extraction, we apply frame difference to obtain motion energy images for subsequent processing.

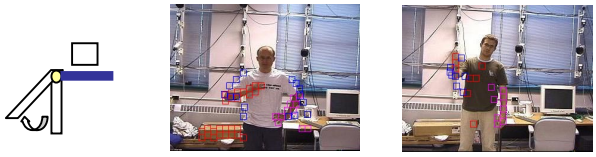
For color feature extraction, we apply face detection algorithm first, and build a skin color model from the detected face region.

4.1.2. Detection of candidate states from features. Candidate states of body joints are needed in belief propagation algorithm. Theoretically, these state can be obtained by uniformly sampling the space of interest, but the potential number of candidate states will make the computational complexity extremely high. Here, we use a more practical approach by first find candidate positions of body joints using the extracted image features. This approach improve the speed by sacrificing some accuracy.

Candidate positions for hands and wrists are detected based on the color model obtained from face detection. Some results are shown in Fig. 6(b). Distracters in background comprise some of the candidate positions, but those are expected.

For elbow and shoulder joints, we use another strategy. We first generate approximate positions of shoulder joints based on assumption of human pose, then we use motion feature to generate candidate positions of elbows. Fig. 6(a) shows the method used in generating the candidate elbow positions. We accumulate motion cues in rectangles with width approximate the width of upper arm, rotating around assumed shoulder joints. For each rectangle at a rotation angle, we cluster motion cues within the rectangle, and find the major connected component of motion cues. The border of the connect component at far end from the shoulder joint is detected as a candidate elbow joint position. After inference in the Markov network, we use motion feature to optimize positions of shoulders, based on estimated position of elbows. Fig. 6(b) gives results of elbow and wrist joints detection.

It is worth noting that the candidate position detection step discussed in this section is used for speeding up the algorithm, and not required by the proposed Markov network model. We can always sampling the space to get the candidate states.



(a) (b)

Figure 6. (a) Detection of elbow joint positions. (b) Detected candidate elbow (blue) and wrist/hand (red and pink) joint positions overlaid on frame images.

4.2. Results

We tested our algorithm on captured videos with people performing meaningful gestures. The videos are recorded with 5 people, each 5 to 10 minutes. We also tested our system on cooking shows and some surveillance videos.

Belief propagation is straightforward to apply in the multi-frame Markov network. In our experiments belief propagation algorithm always converges in several iterations even though the Markov network contains loops.

Fig.7 shows a motion inference result. The estimated 2D joint positions and recovered 3D configuration are given in Fig. 7(a) and Fig. 7(b). By incorporating temporal constraints, the multi-frame model avoids many problems that would be detection errors based on single frame algorithm. Fig.7(d) shows an example of detection error based on a single frame.

Fig.7(c) shows the convergence of beliefs of state nodes after each BP iteration. In this experiment we use 4 state nodes for each frame, and a 7-frames window (total of 28 state nodes) to infer body joint positions. We show the beliefs for all candidate states in 2 state nodes.

In our experiments, we found error in about 12% of the total frames of videos under test. This does not include the cases where the estimation is roughly correct but inaccurate. Errors occur mostly in occlusion situations (Fig. 8) or more subtle situations, such as when two hands are too close together.

Fig. 8 gives an example of inference error caused by occlusion, partly due to the motion energy feature we used. The feature has limited discriminating ability in occlusion situations. We are now working on adding more features to the system in order to deal with some difficult situations.

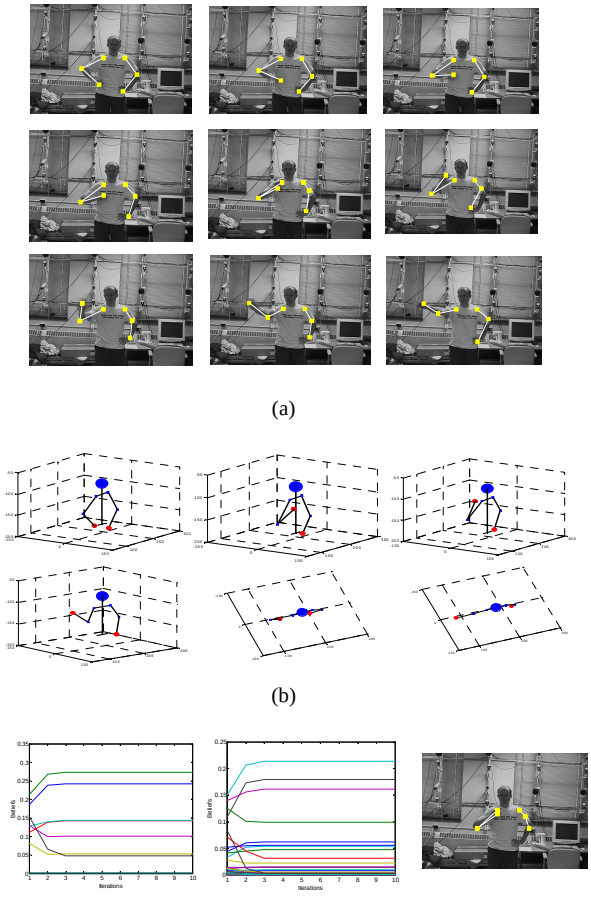


Figure 7. (a) Motion inference result. (b) Recovered 3D stick figures. (c) Convergence of beliefs of 2 state nodes with 9 and 88 candidate states, respectively. Beliefs in 10 BP iterations are shown. (d) A single frame optimal estimation which was corrected by multi-frame constraints in (a).

Figure 8. An error caused by occlusion.

5. Conclusions

Human motion inference and body parts labeling is a difficult problem. So far no existing feature is confidential enough for inference. We believe to solve the problem we have to take advantage of an effective statistical inference approach and a combination of different features.

This paper is an attempt in this direction. We propose a Markov network model for inference of human upper body motion. We utilize belief propagation algorithm for inference in this Markov network. The multi-frame Bayesian inference algorithm using BP give promising results.

In the future, we will improve the algorithm in the following aspects. First, we will compare the results of using detected candidate states and uniformly sampled candidate states. Second, we will utilize more features or better way to use these features, in order to deal with some difficult situations. Finally, find a better solution to the view-specific problem.

References

[1] A. Bobick and J. Davis, The recognition of human movement using temporal templates. IEEE Trans. PAMI, vol. 23, no.4, 2001.

[2] C. Bregler and J. Malik. Tracking people with twists and exponential maps. Proc. IEEE CVPR, pp. 8-15, 1998.

[3] P.F.Felzenszwalb. Object recognition with pictorial structures. MIT AI Technical Report 2001-002, 2001.

[4] W.T. Freeman, E.C. Pasztor, and O.T. Carmichael. Learning Low-Level Vision, *International Journal of Computer Vision*, Vol 40, no.1, pp. 24-57, October 2000.

[5] D. Gavrila and L.S. Davis. 3-D model-based tracking of human in action: a multi-view approach. Proc. IEEE CVPR, pp. 73-80, 1996.

[6] I. Haritaoglu, D. Harwood, and L.S. Davis. W4: Who, when, where, what: a real time system for detecting and tracking people. 3rd IEEE Int. Conf. Automatic Face and Gesture Recognition, Nara, Japan, 1998.

[7] S. Ioffe and D.A. Forsyth. Human tracking with mixtures of trees. P*roc. IEEE Int. Conf. on Computer Vision*, pp. 690-695, July, 2001

[8] M.I. Jordan ed. Learning in Graphical Models, Cambridge: MIT Press, 1998.

[9] S.M. Ju, M.J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. Int. Conf. On Automatic Face and Gesture Recognition, pp. 38-44, 1996.

[10] G. Mori and J. Malik. Estimating human body configurations using shape context matching. Proc. ECCV, pp. 666-680, 2002.

[11] C.J. Taylor. Reconstruction of articulated objects from point correspondences in a single image. Computer Vision and Image Understanding, vol.80, no.3, pp.349-363, 2000.

[12] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Generalized Belief Propagation, *Advances in Neural Information Processing Systems (NIPS)*, Vol 13, pp. 689-695,2001.