Posture and Gesture Recognition using 3D Body Shapes Decomposition

Chi-Wei Chu, Isaac COHEN
chuc@usc.edu, icohen@usc.edu
Institute for Robotics and Intelligent Systems
Integrated Media Systems Center
University of Southern California

Abstract

We present a method for describing arbitrary human posture as a combination of basic postures. This decomposition allows for recognition of a larger number of postures and gestures from a small set of elementary postures called atoms. We propose a modified version of the matching pursuit algorithm for decomposing an arbitrary input posture into a linear combination of primary and secondary atoms. These atoms are represented through their shape descriptor inferred from the 3D visual-hull of the human body posture. Using an atom-based description of postures increases tremendously the set of recognizable postures while reducing the required training data set. A gesture recognition system based on the atom decomposition and Hidden Markov Model (HMM) is also described. Instead of representing gestures as HMM transition of postures, we separate the description of gestures as two HMMs, each describing the transition of Primary/Secondary atoms; thus greatly reducing the size of state space of HMM. We illustrate the proposed approach for posture and gesture recognition method on a set of video streams captured by four synchronous cameras.

1 Introduction

In everyday life, human uses multimodal interactions like speech, gesture or gaze to communicate with each other. Thus a human-computer interaction (HCI) system that incorporates those elements can allow people interact with computers in a more intuitive way. A seamless multimodal user interface is an important element of various immersive systems such as virtual reality, video game consoles or robotics, etc. In these environments, users should be able to interact with computers naturally as they do with human counterparts, or manipulate virtual entities as real world objects. To develop such interaction systems, computers must be able to automatically perceive and identify users' communicative actions such as postures and gestures and responds to them accordingly. The focus of many gesture recognition work today is on using passive sensors such as cameras to capture human actions. The main objective is to construct a vision-based user interface in a natural immersive interactive environment, in which the state and the action of the user(s) can be automatically inferred from a set of video cameras.

Understanding human gestures in an environment by visual perception is a challenging task. A rich data description is required in order to represent both the global and local properties of the perceived features. Furthermore, the dictionary of recognizable postures grows very fast, as we increase the number of desired postures. In this paper we present a method for capturing and describing human body 3D shape for human posture recognition. We show that decomposing the perceived postures into set of elementary postures called *atoms*, allows to recognize a large number of postures and their temporal evolution (*i.e.* gestures) from a very small set of atoms. By decomposing posture descriptors into atoms, we can drastically increase the size of posture dictionary, since we need only to store the atom postures instead of enumerating all possible combinations. Posture recognition is a first step towards gesture recognition. A performed gesture can be viewed as a set of temporal transition of "basic" postures. We present in this paper gesture recognition formalism based on a HMM where the states are the primary/secondary atoms. This HMM formalism enables the recognition of a larger number of dictionary from a small set of learned transitions.

1.1 Previous works

Various methods have been proposed for the recognition and classification of human body postures. Some approaches recognize the postures directly from 2D images. They either fit body models (3D or 2D) into images [2][3][5], or classify postures by image features [1][4]. The main difficulties of estimating human posture directly from 2D images are from the lost information caused by self-occlusion and image projection. Roughly one third of the degrees of freedom of the human model are unobservable due to motion ambiguities and self-occlusions. To compensate for these ambiguities due to 2D acquisition, several approaches rely on using multiple cameras. These approaches rely on two, to an array of cameras or 3D body scanners to capture the 3D human shape and motion. Some of those approaches extract 2D image features from each camera and use these features to

0-7695-2372-2/05/$20.00 (c) 2005 IEEE

search for, or update the configuration of a 3D body model [6][7][8][9]. Others introduce an intermediate step of reconstructing the 3D shape of human body. The characterization of the human pose is then done by fitting a kinematics model into the 3D shape information [10][11], or by using a shape descriptor for classification [12]. The articulated model based approaches provide accurate estimation of body joints configurations, but requires intensive computing and state of the art techniques still lack robustness and accuracy for rapid limbs' motion, as observed during human to human interactions.

1.2 Outline of our Approach

We present in this paper a method for posture identification technique that decomposes the posture into the combination of *elementary postures*, called *atoms*. Our approach is based on the 3D visual hull created from the integration of 2D silhouettes captured by two or more cameras. The surface points are sampled and encoded by a shape descriptor defined by a distribution. The shape descriptor is invariant to people's body proportion (*i.e.* scale), translation and rotation. Continuity properties are also satisfied providing a robust shape descriptor that exhibits localized variation of the distribution for localized 3D shape variation. Moreover, the 3D shape descriptor we propose can selectively encode privileged axis of symmetry or desired rotation invariance. These properties are important for human posture identification, since the human body possesses such a symmetry axis.

Human body posture recognition is challenging task, since it must account for all possible postures to be recognized, as well as posture variations across people. Furthermore, limbs positions and configurations create a large number of postures that translate into small variations on the body shape. This requires a large training data set, and furthermore the classification accuracy decreases as the number of similar clusters increases. Also, time complexity depends on the number of postures available in the dictionary. We propose to address this problem by considering a small set of elementary postures that can describe any posture in the considered dictionary, as a linear combination. Instead of enumerating the sheer number of all possible postures, we select a small set of basic postures called *atoms*. For a given input posture descriptor to be identified, we decompose it using a set of known atoms. The descriptor is then represented as a weighted combination of atoms. This could be used to describe the posture in itself, or used in other classification methods such as SVM.

Parsing continuous variations of postures into the corresponding gestures can also be accomplished by identifying transitions between key postures. We propose a HMM-based gestures recognition method relying on the representation of intermediate postures by a set of primary/secondary atoms. By modeling the transition and observation of atoms instead of each individual postures, the state space and the computational complexity of the HMM can be drastically reduced. An overview of the proposed approach is illustrated in **Figure 1**.

The paper is organized as follows; section 2 describes the 3D shape descriptor and its additive property. Section 3 presents the body postures decomposition and classification inspired from the Matching Pursuit algorithm. Section 4 presents a gesture recognition method based on our primary/secondary posture description and recognition technique. Section 5 shows the experimental result of this work. Then the paper concludes by discussing the results, potential improvements and future work.

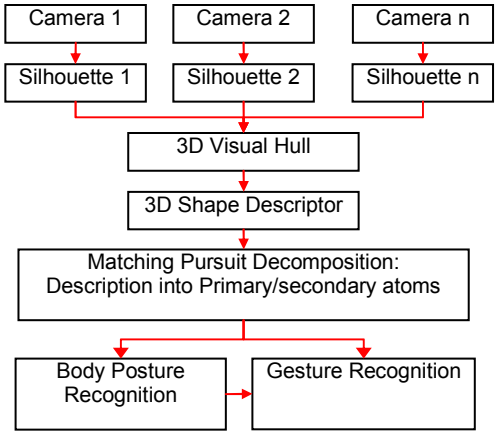


Figure 1: *Overview of the proposed approach*

2 Human Body Shape Descriptor

We use the shape descriptor proposed by [13]. This statistical shape descriptor model preserves the localization of the geometric features considered, while small variations or noise in the shape will induce only small changes in the description and will not interfere with global representation. These properties are crucial for posture recognition.

2.1 3D Human Body Reconstruction

The shape descriptor is computed from the 3D human shape surface. We reconstruct the surface by computing the visual hull from image silhouettes captured from multiple angles. The silhouettes of human body are segmented by a Gaussian background model. The average and variation of each pixel of the background is inferred during the learning phase. Pixels in the new images that exceed certain variation threshold from the average value are considered as new objects in the scene. However, this method will segment real objects from the scene as well as their shadows. We assume that in indoor environment, shadows cast by diffuse light will have blurry boundaries. Therefore edge properties can be incorporated to eliminate shadow regions. Given a set of 2D silhouette images of human body from different angles, we can approximate

the 3D shape of the original object by reconstructing the visual hull [14]. We compute the polyhedral approximation of the visual hull from polygonal representation of the silhouettes [15]. This method is fast and allows us to achieve real-time reconstruction. An illustration of the cameras' configuration and the visual hull is shown in **Figure 2**.

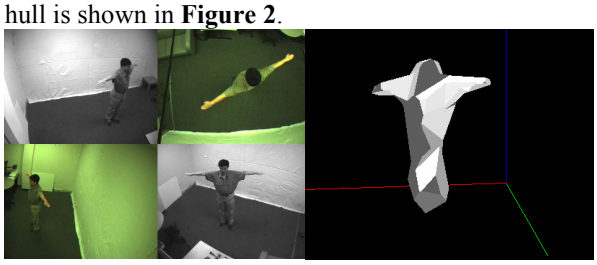


Figure 2: *3D visual-hull reconstructed from 4 images*

2.2 Shape Descriptor

The number of vertices of the polyhedral visual hull depends highly on the polygonal approximation of the silhouette and often is not uniformly distributed on the surface. So we uniformly sample points within polygon triangles of visual hull. We define the shape of a object as the set of sampled surface 3D points $P=\{P_i \mid i=1...N\}$. We compute a bounding reference shape C_R that bounds the visual hull and centered at centroid of the point cloud. In our approach, we use a cylindrical shape, as depicted in **Figure 3**, however other shapes such as spheres could be also used. A set of uniformly sampled reference points $\{Q_j \mid j=1...M\}$ on the reference cylinder are defined. A coordinate system is defined for each reference point: it is centered on the point and tangent to the reference cylinder. For each point P_i on the visual hull and reference point Q_j, We compute the relative coordinate P_iQ_j. This relative coordinate is encoded in spherical coordinate system. That is, $P_i - Q_j = (r, \theta, \varphi)$. The radius r is normalized to [0, 1] with respect to the size of the cylinder. For each reference point Q_j, we construct a K×L×P binned spherical distribution. Each bin $(r_k, \theta_l, \varphi_p)$ stores the number of silhouette points $N_i(k,l,p)$ projected onto that bin. The histograms over all reference points are then summed:

$$N'(k,l,p)=\sum_i N_i(k,l,p)$$

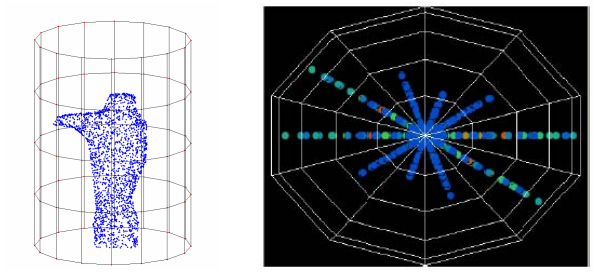


Figure 3: *Left: visual hull and cylindrical reference points selected. Right: the bin distribution plotted in Euclidean space.*

The bin values are then normalized with respect to the largest value:

$$N(k,l,p)=\frac{N'(k,l,p)}{\max_{k,l,p}(N'(k,l,p))}$$

The descriptor of a shape P, $Desc(P)$, is represented as a vector of K×L×P dimensions, recording the normalized value of the bins. The derived shape descriptor is invariant to the scale of the visual hull as the descriptor is normalized by the size of the reference cylinder. It is also translation invariant since the reference cylinder is placed on the centroid. Rotating the posture (and thus the visual hull) around spinal axis is equivalent as a cyclic permutation of the reference points around the cylinder, resulting in an unchanged global descriptor $N(k,l,p)$. Thus the rotation invariance of the shape descriptor is also guaranteed. **Figure 3** depicts such a descriptor. The disadvantage of rotation invariant descriptor, however, is that it cannot distinguish postures between left/right arms. Postures differing only in orientation, such as pointing sideward and forward, cannot be distinguished either.

In this paper we will leverage on another key property of this shape descriptor: its additivity. That is, the descriptor of a composite posture is approximately the additive operation result of the composed sub-postures. Assume given a set of fixed reference points, Q, and two set of surface points S_1 and S_2. Since the unnormalized descriptor records the number of points lying in each bin, then the unnormalized descriptor of the union of the two point sets satisfies:

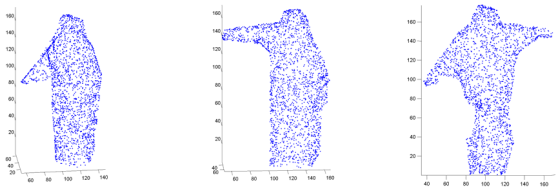
$$Desc(S_1 \cup S_2) = Desc(S_1) + Desc(S_2)$$

This summation property cannot apply to the posture descriptor directly because the sample points set of the composite posture is not the union of the two sets of elementary postures, as they contain overlapping parts such as the torso. Assume we have two postures shape P_1 and P_2, and the posture P_{12} which is the combination of P_1 and P_2, as shown in **Figure 4**. The shape descriptors of these three postures satisfy the following relationship:

$$Desc(P_1) = Desc(\ arm(P_1)\ ) + Desc(\ torso(P_1)\ )$$
$$Desc(P_2) = Desc(\ arm(P_2)\ ) + Desc(\ torso(P_2)\ )$$
$$Desc(P_{12}) = Desc(\ left\_arm(P_{12})\ ) + Desc(\ right\_arm(P_{12})\ ) + Desc(\ torso(P_{12})\ )$$

Since the descriptor is scale and rotation invariant, each body part will have similar descriptor values:

$$Desc(\ torso(P_{12})\ ) \cong Desc(\ torso(P_1)\ ) \cong Desc(\ torso(P_2)\ )$$
$$Desc(\ P_{12}\ ) \cong Desc(\ arm(P_1)) + Desc(\ arm(P_2)) + Desc(\ torso(P_{12})$$
$$\cong Desc(P_1) + Desc(P_2) - Desc(\ torso(P_{12})\ )$$



(a) (b) (c)

Figure 4: *(a) Posture P_1 (b) Posture P_2 (c) Composite posture P_{12}.*

However, the descriptor is a global shape descriptor and does not separate between different body parts. We use a "resting posture" that representing "stand still" posture. And use this to compensate for overlapping torso part

$$Desc(P_{12}) \cong Desc(P_1) + Desc(P_2) - Desc(Resting)$$

Each posture descriptor is implicitly normalized with respect to the number of sample points, so the descriptor values will not biased by the sampling rate. Thus the relationship of the elementary and composite descriptor is more accurately represented by a "bin-occupancy" operator, rather than a summation. Such property will be exploited in the next section to decompose complex postures into set of simple basic postures.

3 Decomposing Shapes

Recognizing arbitrary human body posture is a challenging task as it has to take into account the variability across people in executing the same posture. We have used the shape descriptor directly with data classifiers such as support vector machines (SVM) to recognize different postures. We collected training data for all postures to be recognized, and trained a SVM classifier for each pair of them. New input posture descriptor was fed to those SVMs and the best fitting posture was selected. The advantage of using SVM classifiers is the high tolerance of noise and the recognition accuracy; however this method requires training the SVM on every posture to be accounted for. Using a SVM for each pair of postures results in a large number of SVMs, although a hierarchical classification can be used to reduce the number of SVMs [13].

The additivity property of the shape descriptor, suggests that the recognition of complex body postures can be achieved by recognizing a subset of elementary postures called *atoms*. In this section, we propose a method for decomposing arbitrary input posture as the weighted sum of *atoms*. Such decomposition can be used to represent and recognize a large dictionary of complex postures from a small set of atoms.

3.1 Matching Pursuit Algorithm

A common representation of signals relies on the adaptive approximation technique. Such approach seeks to find the representation of function *f* as a weighted sum of elements from an overcomplete dictionary. Given a signal *f* and a redundant dictionary *D* as a collection of signals $D = \{g_\gamma | \gamma \in \Gamma\}$, this technique seeks to decompose the original signal *f* as a linear combination:

$$f \cong \sum_{\gamma \in \Gamma} \alpha_\gamma g_\gamma, \alpha_\gamma \in \Re$$

The optimal approximation of $[\alpha_\gamma]$ is the one that results in the weighted sum that most closely resemble the original function. Various methods have been proposed to find the optimal decomposition, such as: method of frames [16], best orthogonal basis [17] and basis pursuit [18]. Each method places different constrains on the $[\alpha_\gamma]$ vector, such as minimizing its L1 or L2-norm, and solves the weights accordingly. The original function can thus be represented by series of weight parameters. However, we want to use such approximation for feature selection instead of data compression. We intent each posture *f* to be approximated by only as few elements in the dictionary as possible, and each element closely matches the local properties of *f*. Finding the optimal approximation over a redundant dictionary was proved to be a NP-complete problem [20]. However, the matching pursuit (MP) algorithm proposed in [19] avoids such complexity. The MP algorithm assumes the input signal *f* and the basic elements, called *atoms*, in the dictionary $D = \{A_i | i = 1 \ldots M\}$ are all in the Hilbert space. It also assumes that all atoms are normalized to unit length. The MP algorithm is iterative: at each iteration, it computes the n-th residue R_n, starting with $R_0 = f$, the MP algorithm chooses the atom $A_i^n \in D$ that maximizes the absolute value of its scalar product with residue left from previous iteration (the projection of R_{n-1} onto A_i^n), which is equivalent to minimize the magnitude of the next residue $\|R_n\|$. The atom $A_i^n \in D$ is defined by:

$$A_i^n = \arg\max_{g_i \in D} |\langle R_{n-1}, A_i \rangle|$$

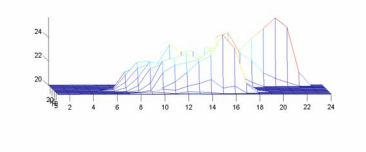
where $\langle \cdot, \cdot \rangle$ is the scalar product operator, and R_{n-1} is defined by: $R_n = R_{n-1} - \langle R_{n-1}, A_i^n \rangle A_i^n$

The main advantage of MP algorithm compared to other methods is its efficient computation: instead of solving for a global optimization, MP uses a non-optimal greedy method in each step, and chooses the element that reduces the most the residue function.

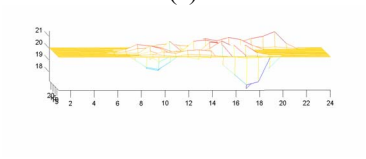
3.2 Decomposing Postures

Posture data, however, has one major difference compared to atoms used in signal or image decomposition: all the posture atoms have a large overlapping part at the torso, legs and head section. The densities of bins around torso are usually much higher than densities corresponding to arms and hands. This makes the original matching pursuit unstable for use in the posture decomposition process. Indeed, after the first iterations, bins at torso part of the residue will have large negative values due to repeated subtraction, and subsequent iterations will then focus on compensating the negative bins instead of trying to fit the actual arms and hands features as depicted in **Figure 5**.

We propose a modified matching pursuit algorithm suitable for decomposing a posture *f* in to a set of atoms. Assuming we have a dictionary of postures *D={A_i | i = 0 to M},* where the first atom A_0 is the resting posture, *i.e.* the common denominator of all postures. The input posture *f*, and the atoms are normalized to unit length. Then the modified MP decomposition estimates the residues R_n, starting with:



(a)



(b)

Figure 5: *Descriptor bin densities plotted as mesh. x, y axis are r and φ bin indices, θ is fixed as zero. (a) The original posture descriptor. (b) The residue after first two iterations. There are large negative sections in the residue.*

$$R_0 = f - \langle f, A_0 \rangle A_0$$

and chooses the atom $A_i^n \in D$ such that:

$$A_i^n = \arg\max_{g_i \in D} (overlap(R_{n-1}, A_i))$$

where the residues R_n, are defined by:

$$R_n = overlap\_diff(R_{n-1}, A_i^n)$$

The *overlap(f, g)* function returns the number of non-empty bins where *f* and *g* overlap. In each iteration, the algorithm will select the atom that has the most overlapping bins with previous residue. And the *overlap_diff(f,g)* function removes from *f* the overlap bins among *f* and *g*.

By removing the resting posture A_0 before applying matching pursuit, we prevent other atoms from absorbing the torso values, which are considered as non-features they do not characterize the variations among atoms. This will keep the weights of the atoms balanced. We also operate only on the overlapping property instead of the bins' density. This ensures that any noise on the torso region that was not subtracted in the first step will not interfere with the atom selection process in the subsequent iterations as in the original MP algorithm. After *M* iterations, we compute the weight of each atom as follows: $\alpha_0 = \langle f, A_0 \rangle$ and:

$$\alpha_i = \sum_{n, A_i^n = A_i} \langle R_{n-1}, A_i^n \rangle, i = 1 \ldots M$$

This will result in a weight vector $\{\alpha_i \mid i=0,1 \ldots M\}$, which is far smaller than the dimension of the descriptor.

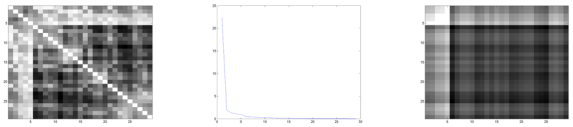
After estimating the weights of the decomposition of the posture *f* into the set of selected atoms A_i, it is straightforward to recognize the composition of the postures. First, a *posture threshold* is applied to the weights to eliminate atoms decomposed from noise. Then the two atom postures, different of A_0, with largest weights are selected and constitutes *primary* and *secondary* atom postures, ordered by their lexical precedence. Note that the resting posture A_0 is always present and has largest weight, as it absorbs the bins at the torso part. It is possible that after thresholding, none or only one atom is left. Such posture will be classified as resting posture or the corresponding atom A_0.

Frequently, the posture *f* corresponds to multiple instances of the same atom. For example, the posture representing two arms up in symmetric configuration, corresponds to twice the contribution of the atom representing one arm. The proposed matching pursuit algorithm identifies automatically such situations by analyzing the estimated weights. By the additive property, the densities of the arms part of symmetric postures are about twice than the densities of corresponding one-arm posture. And these higher densities will reflect on the value of weights that are computed from the scalar product of densities vectors. To solve the multiple-instances problem, we collect training data for each one-atom elementary posture and decompose them by the corresponding atom, and record the average weight α_i'. After decomposing an arbitrary posture, we compare the weight of each atom, α_i, to the corresponding α_i'. If the ratio of the two weights exceeds certain *instance threshold*, that atom is considered to be of multiple instances and the primary and secondary atoms are marked to be the same atom.

3.3 Selection of Atoms and Learning

An essential element in the posture decomposition is the selection of atom postures. The atoms in the posture dictionary must be discriminative; otherwise similar atoms will compete with each other in the matching pursuit process, resulting in low weight value distributed over multiple atoms. The choice of the histogram bin resolution also affects the discriminative power of the atoms.

To select the most distinctive atoms, we collected 30 different arbitrary postures. For each pair of postures P_i and P_j, we ran the matching pursuit algorithm, using P_j as the atom to decompose P_i, and record the resulting weights in a 30x30 symmetric matrix *M* shown in **Figure 6**. This matrix is then decomposed by Singular Value Decomposition (SVD) and its lower rank approximation is computed. From the lower ranked matrix we selected five atoms corresponding to the largest eigenvalues, these allow to extract the corresponding elementary postures. These atoms, depicted in **Figure 7**, will serve as atoms in the posture decomposition process. This selected set of postures generates a dictionary of a total C(5,2)+5 = 15 recognizable composite and elementary postures.



(a) (b) (c)

Figure 6 :*(a) The resulting weight matrix M of matching pursuit decomposition of each pairs of 30 postures. (b) Singular values of M. (c) Rank 1 approximation of M*

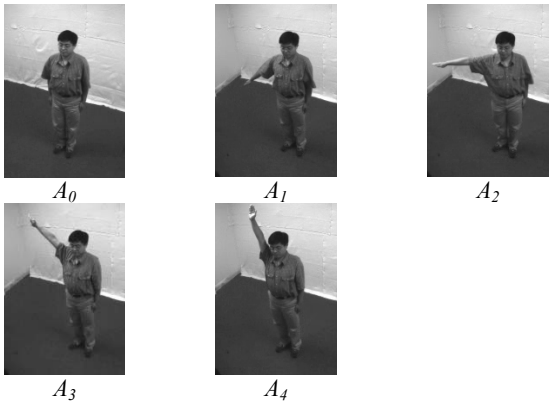


Figure 7: *Five atom postures selected automatically by the SVD analysis.*

4 Gesture Recognition: Dual-state Hidden Markov Model

The posture recognition method presented in this paper provides an efficient description for gestures analysis. It parses continuous variations of the postures into occurrences of the elementary postures or atoms available in the dictionary. In this section we will present a formulation of a Hidden Markov Model (HMM) relying on the primary/secondary decomposition of arbitrary postures for gestures recognition.

The elements of a HMM consists of a set of states $S = \{s_1, s_2, \ldots s_N\}$, a set of observation symbols: $O = \{o_1, o_2, \ldots o_M\}$, a state transition probabilities matrix $A = \{a_{ij}\}$, $a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i)$, $1 \le i, j \le N$, and a state-observation probabilities matrix: $B = \{b_{jk}\}$, $b_{jk} = P(o_k \mid s_j)$, $1 \le j \le N$, $1 \le k \le M$ and finally, the probabilities of initial states:

$$\pi = \{\pi_i\},\ \pi_i = P(q_1 = s_i),\ 1 \le i \le N$$

Given the elements above, a HMM is often represented as (A, B, π). Since [21], HMM have been widely used by researchers to model and recognize the temporal or spatial transition of gestures [22][23]. In proposed methods for gesture recognition, each gesture is defined as a HMM with different state transition matrix A, and each state is defined as a posture or associated to a motion profile. The aforementioned methods seek to find the most probable model among available gestures that describes the observed posture sequence. These approaches, however, require an extremely large number of states space as the number of gestures to be modeled increases. To address this limitation, we propose a formulation of HMM based on *atoms* instead of complete postures. That is, we have state space is defined by the atoms: $S = \{A_0, A_1, \ldots A_N\}$ and the set of observations that corresponds to the decomposition of each posture into a primary and secondary atom: $O = \left\{\left(A_i^p, A_j^s\right)\right\}$, $0 \le i,j \le N$.

The set of observations represents all possible decomposed pairs of primary/secondary atoms. The transition matrix A and observation matrix B are also defined based on atoms, respectively. $A = \{a_{ij}\}$, $a_{ij} = P(q_{t+1} = A_j \mid q_t = A_i)$, $0 \le i, j \le N$, and $B = \{b_{jk}\}$, $b_{jk} = P(o_k \mid A_j)$, $1 \le j \le N$, $0 \le k \le M$.

In our framework we assume the initial state is always the resting posture: $\pi = \{\pi_i\}$, $\pi_0 = 1$, $\pi_i = 0$, $1 \le j \le N$. However, this practical assumption does not reduce the scope of the proposed approach, since we can easily detect the resting posture.

In this paper we assume that the transition and observation of the Primary/Secondary atoms in a gesture are *independent*. Assume G is the dictionary of all pre-defined gestures. Then each gesture in G consists of two HMMs instead of one, one for the primary atom and the other for the secondary atom, each with different transition matrices. The considered HMM is then defined by: $G = \{m_i\}, m_i = \left(m_i^p, m_i^s\right), i = 1 \ldots \|G\|$,

and $m_i^p = \left(A_i^P, B, \pi\right)$, $m_i^S = \left(A_i^S, B, \pi\right)$

The input sequence of K postures is decomposed into a sequence of primary/secondary atom compositions $d = \left(v_1, v_2 \ldots v_T\right)$, where $v_t = \left(A_t^p, A_t^s\right)$, $d^p = \left(A_1^p, A_2^p \ldots A_T^p\right)$, $d^s = \left(A_1^s, A_2^s \ldots A_T^s\right)$. For every input vector d, we want to find the most probable gesture in the dictionary by solving:

$$m = \arg\max_{m_i \in G} p\left(m_i \middle| d\right)$$

Assuming that the prior probabilities of all gestures are equal, the problem becomes finding the gesture model that has the highest likelihood for a given observation.

$$m = \arg\max_{m_i \in G} p\left(m_i \middle| d\right) = \arg\max_{m_i \in G} p\left(d \middle| m_i\right)$$

And the likelihood is defined as:

$$p\left(d \middle| m_i\right) = p\left(d^p \middle| m_i^p\right) \times p\left(d^s \middle| m_i^s\right)$$

The probability can be computed by forward algorithm.

Parsing gestures into respective atoms, we reduce the original HMM model with possible $O(N^2)$ size of state space into two HMM models, each with $O(N)$ size state spaces.

5 Experiment Results

We have used the proposed 3D shape descriptor and posture decomposition technique for identifying user's postures while performing specific gestures. The experimental environment consists of four synchronized cameras, allowing real-time image extraction, silhouette segmentation, and 3D human body visual hull reconstruction at 12 frames per second. We used a cylindrical reference shape of 5 vertical by 16 horizontal reference points to infer the shape descriptor. We chose the bin resolution to be $24(r) \times 24(\theta) \times 24(\varphi)$. This resolution allows us to select 5 atoms as depicted in section 3.3. Higher resolution can capture greater local details of the posture shapes, allowing more atoms, but also requires more computational resources.

5.1 Posture Recognition Results

From the collection of 30 different postures, we have used the 5 postures including the resting posture, labeled as A_0 to A_4, as atoms. These postures were selected by the SVD analysis as described in section 3.3. These atoms will be used in the remaining for characterization of postures and gestures. We select a posture threshold value of 0.0042, all atoms' weight below this value will be discarded as noise. The instance threshold is 1.4. We collected 1000+ samples of each single atom posture and trained the average weight of the corresponding atom after decompositions. If after decomposition, the ratio between an atom weight and its average weight exceeds this threshold, the atom is considered of multiple instances. The proposed approach allows recognition of postures across people without a user-specific training data set. To show the recognition rate across persons, we collected video sequences of all possible atom combinations; each contains about 1000 frames of two different persons. The average recognition rates of different postures are shown in **Figure 8**. The atoms were trained on the first person (left)

Person / Postures			
A_0	100	100	94.6
A_1	88.4	88.7	80.9
A_2	85.8	91.0	87.7
A_3	91.9	82.7	97.0
A_4	81.6	94.2	79.4
$A_1 A_1$	97.2	98.5	96.2
$A_1 A_2$	95.1	85.8	86.8
$A_1 A_3$	94.0	74.4	90.8
$A_1 A_4$	90.5	73.1	72.3
$A_2 A_2$	99.2	91.5	89.9
$A_2 A_3$	99.1	98.1	88.2
$A_2 A_4$	94.3	70.2	87.9
$A_3 A_3$	85.0	70.8	72.4
$A_3 A_4$	99.7	94.6	81.7
$A_4 A_4$	88.6	74.3	66.9

Figure 8: *Recognition rate of different composite postures performed by different users. The model has been trained on training data from the person on the left. The training and testing data sets are different for the first person.*

5.2 Effect of Decomposition Parameters

Choosing the posture and instance thresholds value is the tricky part: a too small posture threshold, then minor noise surface points on the visual hull result from image segmentation error or self-occlusion of human body will be classified as atom postures. Setting the threshold too high, then many valid decomposition will be filtered out, resulting composite posture been classified as basic

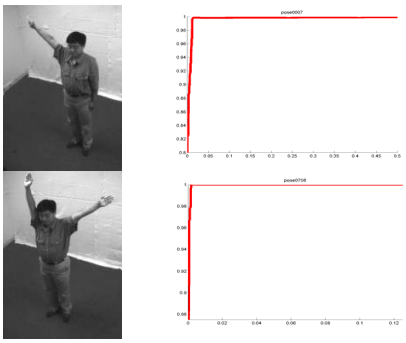


Figure 9: *ROC Curves of the decomposition of an atom posture and of a composite posture using the proposed Matching Pursuit algorithm.*

postures or even resting postures.

The effect of different threshold and the discrimination power of our recognition method can be estimated by plotting the Receiver Operating Characteristic curve (ROC curve). To evaluate the MP algorithm, we collected 1000 data sample for each possible composite and atom posture. Then apply the MP decomposition and recognition for each posture. Each posture records its rate of true positive (correctly recognized) and false positive cases under different thresholds, and plot on the ROC curve. As shown in the example ROC curves in **Figure 9**, the curves are tightly bounded to left and top boundaries, indicating that our posture recognition method is very accurate.

5.3 Gesture Recognition Results

The second set of experiments focused on assessing the performance of the primary/secondary HMM approach for gesture recognition. Due to time limitation, we defined six simple gestures by hand-crafting the primary/secondary atom state transition matrices instead of training them. And we trained the observation matrix by over 10000+ samples of all atom combinations. The gesture testing data are collected from two subjects. Each person performed pre-defined gestures with different rotation variances, such as pointing forward or sideward, as shown in **Figure 10**. Since our posture descriptor is rotation invariant, both forward and sideward pointing actions were correctly

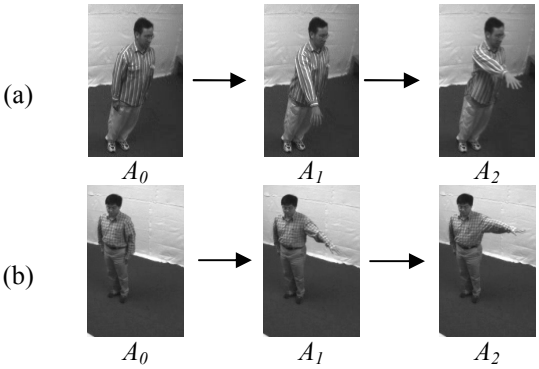


Figure 10: *Two examples of a simple gesture (a) Pointing forward. (b) Pointing sideward*

recognized as the same "pointing" gesture. More complex gestures modeled by this dual-HMM can also be correctly classified (**Figure 11**). However, since we assume that the primary/secondary state transitions are independent, gestures that only differ at the synchronicity of the two atoms cannot be distinguished. For example, two arms point forward at the same time or one arm points after the other are considered the same gesture. Thus special care must be taken to design the primary/secondary state transition matrices of those gestures.

6 Conclusion

Identifying user postures is a first step towards the challenging task of gestures recognition. In this paper we presented a method to characterize postures as a composition of basic primary and secondary atoms directly from their 3D shape descriptor. For this purpose we have reformulated the matching pursuit algorithm. The description of an arbitrary posture into a primary and secondary atom provides compact representation of a large dictionary of postures using a small dictionary of atoms. The proposed method allows for recognizing a large set of postures from a small set of atoms.

Similarly, the formulation of a HMM relying on the primary/secondary decomposition provides a more efficient approach for gestures recognition by allowing a larger descriptive power using a small set of atoms. We have investigated the characterization of basic gestures, or gestemes as a transition states models of the canonical body posture atoms. Our experimental classification of the posture transitions are very encouraging and indicate a strong temporal structure in the primary/secondary atoms that could be used for robust gesture inference.

Acknowledgments

This research was partially funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, under Cooperative Agreement No. EEC-9529152.

7 References

[1] A. Elgammal, C.S. Lee "Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning" IEEE CVPR 04, Washington, DC, June 26-July 2nd, 2004.

[2] Z. Chen, H-J. Lee, "Knowledge-guided visual perception of 3-D human gait from a single image sequence" Systems, Man and Cybernetics, IEEE Transactions on, Vol.22, Iss.2, Mar/Apr 1992 Pages:336-342

[3] P. Horain, M. Bomb, "3D model based gesture acquisition using a single camera" WACV 02. Vol., I., 2002 Pages: 158- 162

[4] M.M. Rahman, S. Ishikawa, S. "Appearance-based representation and recognition of human motions" ICRA 03. Vol.1, Iss., 14-19 Sept. 2003 Pages: 1410- 1415 vol.1

[5] Parameswaran, V.; Chellappa, R. "View independent human body pose estimation from a single perspective image", CVPR 04, Vol.2, Iss., 27 June - 2 July 2004 Pages: 16- 22

[6] T. Izo, W.E.L. Grimson, "Simultaneous Pose Estimation and Camera Calibration from Multiple Views", IEEE Worskhop on Articulated and Nonrigid Motion at CVPR, June 2004.

[7] J. Deutscher, A. Blake, and I. Reid. "Articulated body motion capture by annealed particle filtering". CVPR 00, volume 2, pages 126–133, 2000.

[8] Changbo Hu; Qingfeng Yu; Yi Li; Songde Ma. "Extraction of parametric human model for posture recognition using genetic algorithm" Automatic Face and Gesture Recognition, 2000. Vol., Iss., 2000 Pages:518-523

[9] H. Yoshimoto, N. Date, S. Yonemoto. "Vision-based real-time motion capture system using multiple cameras" MFI 03, Vol., Iss., 30 July-1 Aug. 2003 Pages: 247- 251

[10] C.W. Chu. O.C. Jenkins, M.J. Mataric', "Markerless kinematic model and motion capture from volume sequences", CVPR 03. Volume: 2 , 18-20 June 2003 Pages:II-475 - II-482

[11] I. Miki´c, M. Trivedi, E. Hunter, and P. Cosman. "Articulated body posture estimation from multi-camera voxel data", CVPR 01, Kauai, Hawaii, December 2001.

[12] N. Werghi, Yijun Xiao. "Posture recognition and segmentation from 3D human body scans". 3D Data Processing Visualization and Transmission, 2002. 636- 639

[13] I. Cohen and H. Li, "Inference of Human Postures by Classification of 3D Human Body Shape" ICCV'03.

[14] A. Laurentini, "The visual hull concept for silhouette-based image understanding", IEEE Trans. on Pattern Analysis and Machine Intelligence, 16(2):150–162, 94.

[15] W. Matusik, C. Buehler, L. McMillan. "Polyhedral Visual Hulls for Real-Time Rendering", Proceedings of Eurographics Workshop on Rendering 2001.

[16] I. Daubechies, "Time-frequency localization operators: a geometric phase space approach", IEEE Trans. Inf. Theory, 34 (4), pp. 605--612, 1988.

[17] R. R. Coifman, M. V. Wickerhauser, "Entropy-based algorithms for best basis selection", Information Theory, IEEE Transactions on, Vol.38, Iss.2, Mar 1992,713-718

[18] Shaobing S. Chen, David L. Donoho, "Atomic decomposition by basis pursuit", Tech. report, Stanford University, Statistics Department, 1995.

[19] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries IEEE Trans. on Signal Proc.,12(41), 3397-3415, 1993.

[20] B. K. Natarajan, "Sparse Approximate Solutions to Linear Systems", SIAM J. Computing, vol24 (2), pp. 227-234

[21] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," CVPR, 1992

[22] G. Ye, Jason J. Corso, G. D. Hager "Gesture Recognition Using 3D Appearance and Motion Features" CVPR 2004

[23] Y. Nam, K. Wohn, "Recognition of space-time hand-gestures using hidden markov model," ACM Symposium on Virtual Reality Software and Technology, 1996, pp. 51--58.

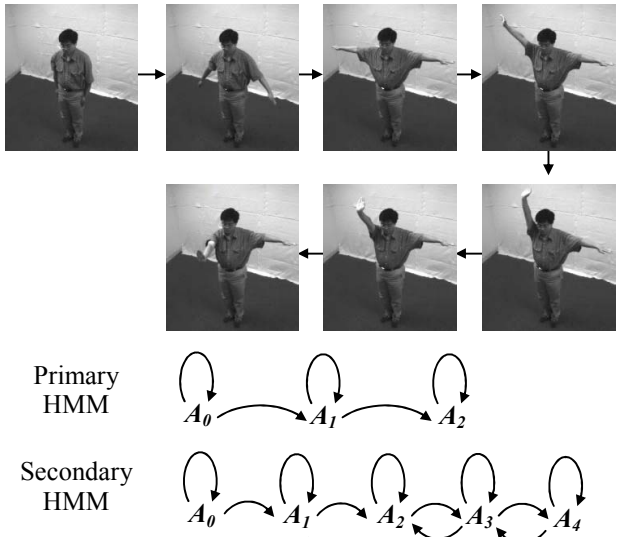


Figure 11: *An example of a complex gesture and its primary/secondary state transition models*