Arm and *Body gesture recognition*[*]

Cédric Graf
Rue de la carrière 9
1700 Fribourg
cedric.graf@unifr.ch

ABSTRACT

This paper presents a survey of different methods which describe the recognition of body and arm gesture. The methods presented in this work cover the topic of tracking and detection of arm and body gestures. But it will also present methods of recognition with help of pattern recognition methods.

1. INTRODUCTION

Gesture recognition of the body and arm provides the basis for important applications in computer science. It gives the base for Human Computer Interfaces without interaction of Hardware. But it also supplies the possibility of the interaction with robots as Yang [11] did it. Other fields use the enhancement of information which are provided in gesture to build useful applications. Wilson [8] for example uses gesture to catch the significant part in a discourse. By doing so, he claims to be able to compress a video stream with a minimal loss of information. Glowinski [3] used the energy provided and detected in gestures to find emotional states of persons. Balder [2] used research which gained emotional and intentional informations out of gesture recognition to improve his avatar. The aim of his research is to improve the gesture of an avatar towards natural like gestures.
In face of the applications offered by arm and body recognition, a survey of different methods used in this field seems to impose itself. This document should not be taken as a complete overview of the subject. It merely aims to give an entry point in to the subject. To do so different methods of body and gesture recognition are presented.
By having a glance at these methods we split them into two approaches. One approach used is simply using tracking and detection methods which give the position of the body and arms in a 2D or 3D space. Another approach is to use pattern recognition methods to extract meaningful information out of gestures.
Based on this observation this work will be divided in two major parts. The first part will look into the detection and tracking of gestures and the second part will deal with pattern recognition methods applied to gesture.

[*]This document was written in the context of a seminar of the research group document, image and voice analysis (DIVA) of the University of Fribourg.

2. GESTURE DETECTION AND TRACKING

In this section we treat gesture as the change of position of body parts in relation to time. According to this definition methods which intend to track and detect body and arm gestures have to be able to locate them in a 2D or 3D space. To reflect this definition three summaries of different authors are presented in this section. Each of them uses a different approach to solve this task. The first uses a Gaussian color model, the second one uses differences in covariance matrix and Kahlman filter and the third one uses an integration of the two first one coupled with spacial constraints of the body.

2.1 Detection and tracking by skin color

Waldherr [7] suggests a 2D method of gesture recognition to give orders to robots.
To do so the camera must be able to capture colors in the RGB color space. His method is based on the recognition of human skin color which can be easily extracted from a scene. He builds a Gaussian color model using the Mahalanobis distance $f(r_i, g_i)$ of the chromatic color r and g of the pixels:

$$f(r_i, g_i) = \frac{1}{2\pi \left|\sum\right|^{0.5}} e^{-\frac{(r_i-\mu_r, g_i-\mu_g)^t \sum^{-1} (r_i-\mu_r, g_i-\mu_g)}{2}} \quad (1)$$

Whereas $\sum$ represents the covariance matrix. μ_r and μ_g are the means of r and g.
Since he uses predetermined pattern of the arm to give orders to a robot. He needs the representation of the arm to match this pattern. To do so he uses the color of the shirt detected a few centimeter below the face. By applying the same color distribution model to the color of the shirt he is able to extract the arm from the scene. The computation of this method has to be performed on each frame to track gestures.
The color model alone has a flaw. Natural environment can change in brightness which would make the distribution of the color model obsolete. To encounter this problem he uses a leaky integrator and updates the color distribution model as follows:

$$\sum_{face}^{t} = \alpha \sum_{face}^{*} + (1-\alpha) \sum_{face}^{t-1} \quad (2)$$

$$\mu_{rface}^{t} = \alpha \mu_{rface}^{*} + (1-\alpha)\mu_{r_{face}}^{t-1} \quad (3)$$

$$\mu_{gface}^{t} = \alpha \mu_{gface}^{*} + (1-\alpha)\mu_{g_{face}}^{t-1} \quad (4)$$

Where $\sum_{face}^{*}$, μ_{rface}^{*} and μ_{gface}^{*} denote the values which are obtained from the most recent image.
By testing his method in natural environment he could find two major flaws. At first he points out that the face of the person which is detected by his system has always to be visible. This situation is not always fulfilled in a natural environment. The face could be covered by obstacles. Another flaw is the lack of recognition of an individual person in a crowd.

2.2 Detection and tracking by 2D clusters

Wren [9] uses a RGB camera to represent the body in 2D clusters which are named blobs. The aim of this work is the real-time tracking of the human body to use it as an avatar. To initialize the blob the covariance of the empty scene is computed. As soon as a body is put in the scene the deviation of this covariance matrix is detected. A contour analysis of this deviation enables with help of the Mahalanobis distance as in the previous section to build the mean color distribution μ_k and the covariances $\sum_k$ of each blob k out of their center. The body parts give the number of 2D clusters. The likelihood d_k of each pixel can now be computed to be part of a blob:

$$d_k = -\frac{1}{2}(y-\mu_k)^T \sum_{k}^{-1}(y-\mu_k) - \frac{1}{2} ln \left| \sum_{k} \right| - \frac{1}{2} ln(2\pi) \quad (5)$$

The max. likelihood assigns each pixel to a blob. The blobs are contained in a support map $s(x,y) = argmax_k(d_k(x,y))$. After having generated the blobs we still have to update them in regard of his movement and of the change of brightness.
To correct the influence of brightness, a leaky integrator as in the previous section is used.
To track gesture a Kahlman filter (G), which takes the location of the blob (X) his velocity (Y) and acceleration into account, is used to predict the future location $X_{[n|n]}$ of the blob:

$$X_{[n|n]} = X_{[n|n-1]} + G_{[n]}\{Y_{[n]} - X_{[n|n-1]}\} \quad (6)$$

Since errors can occur in the blob model, the model can be enforced by prior knowledge like the color distribution of the skin.
Several flaws of the system are pointed out. At first the generated clusters degrade slowly with time. The second one is the assumption of the system to have a static background. If dynamic background occurs the system is not working properly. The third flaw is that the system can not work in a crowd.

2.3 Detection and tracking by multiple cues

Azoz [1] developed an application which makes the tracking and localization of the human arm in 3D space possible.

Figure 1: Sampled frames: The output of the framework is superimposed as a stick model on the real arm.

The application uses like in section 2.1 the skin color as a base to find hands and head. By taking the color a few centimeter below the face, he also finds the color of the shirt. The shoulders are found relative to the position of the head. To be able to localize the elbow he uses time varying edges, which is a method based on gradient detection to find the edge of an image. The elbow is located at the edge which has the greatest distance from the shoulder and the hand.
He than builds a geometrical model of the arms. He enhances the precision of the found arm by fitting the found shoulder, elbow and hand in that model. The model enables to gain 3D information of the arms (see figure 1).
He also uses a Kahlmanfilter see section 2.2 to optimize the region in which he is looking for head and hands. The Kahlman filter gives the ability to find hand and head even if they are obscured.
By testing his framework he could points out that the system was working quiet well. The Kahlman filter continues to track the arms even if they are occluded and wrong detected, time varying edges are corrected by the constraint of the arm model. A loss of precision was found as he compared his framework with magnetic trackers.

3. GESTURE CLASSIFICATION

This section presents pattern recognition methods to extract meaningful information out of gestures. The first two subsections will show examples of classification with help of Hidden Markov Models. The third subsection will present a semantic classification tree.

3.1 Classification of informational gesture with help of Hidden Markov Model

In his paper Wilson [8] detects temporal structures in gesture. With help of these temporal structures he extracts gesture which underlines informational significant sequences of a video. He aims to compress the video to this sequences. The compressed video should only contain informative relevant sequences.
To determine significant gestures in human communication Wilson [8] uses the following gestures:

- iconic, where the movement of the hand matches situations or objects of the narration.
- deictic, which is a pointing gesture.
- metaphoric, where the movement of the hand is somehow suggestive of the situation.
- beats, which are used to correct mis-spoken segments.

Each of these gestures begins out of a rest-state. For example a beat gesture begins in a rest-state, makes a short baton-like movement and returns to the rest-state. This movement can be named bi-phasic.
Iconic, deictic and metaphoric movements can be described as tri-phasic. Their movements strat out of a rest-state merge into a gesture position where they remain and after a while they return to the rest-state.
To classify these movements he uses a decomposition of frames in eigenvector like in Matthew's [6] method. The video sequence is in gray scale. To generate these eigenvectors, a mean matrix of all frames is generated. Out of the difference of a frame and the mean matrix the covariance matrix is computed. By taking the eigenvector of the covariance matrix he gains a modified frame, which underlines the movement distribution of the frames. This frame are named eigenfaces.
By computing the euclidean norm of each of this eigenfaces he gains a difference matrix which has the dimension of the number of the frames. Since the brightness of each pixel in the difference matrix match the euclidean norm, we can conclude that bright rows in the matrix indicates long rest states.
He then computes the probabilistic densities of the rest states to use them as training data for the Hidden Markov Model. By applying these model he could extract bi- and tri-phasic movements. Bi-phasic movements begin out of a rest state (R) go to a transition (T) and return to a rest state (R-T-R). Tri-phasic movements go out of a rest state, merge into a transition state, perform a stroke (S) which is a smaller movement in front of a subject, and then go back to a transition, and finally ends in a rest state (R-T-S-T-R).
By using the model he could parse videos to extract informational significant sequences.
To test the model he took test persons and let them tell a story after they past a stressful situation. 40 persons where involved in this experience. The group of persons could be split in two. In the first goupe he could detect rest-states, in the second group his framework was not able to detect any rest-states. In the first case his system worked well. But for person of the second group the method failed completely.

3.2 Classification of body gesture with help of Hidden Markov Model

Yang [11] developed a system capable to detect body gestures such as touching a knee and wrist, rising a right hand, walking, waving a hand, running, sitting on the floor, lying down on the floor, jumping and getting down on the floor. He does so by using a Hidden Markov Model.
At first he uses a pose reconstruction method from Lee [10] to detect human subjects out of video frames and creates a 3D representation of it. To do so 2D frames of a human beeing are taken by different angles. The 2D shapes of the human are then matched with a 3D model of the human body, with help of the least square minimization method. This method allows to allocate each upper shoulder, elbow, wrist, knee, ankle etc. to their position in the 3D space.
In a second step he needs to extract feature vectors. The previously presented 3D model gains for each frame of a video the structural feature points of the body (wrist, elbow, shoulder, etc.). The center of this 3D space is located in the region of the trunk of the 3D human model.

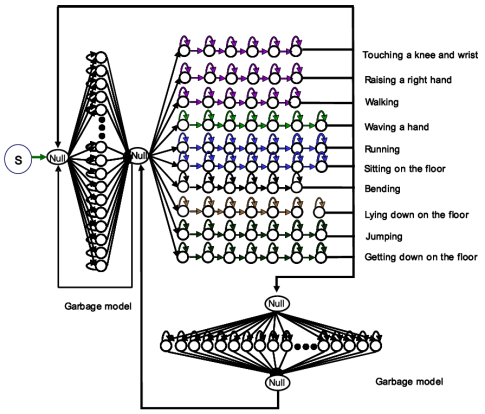

Figure 2: Key gesture spotting model

Each of these points are first projected into the x, y and z plane. Then the angle between this projection points and the axis are measured. These angles give the feature $F_k = (\theta_x, \theta_y, \theta_z)$. For each of this feature we build a feature vector $X_t = [F_{L-shoulder}, F_{L-elbow}, F_{L-Wrist}, ...]$.
In the last step a Hidden Markov Model is used to extract touching a knee and wrist, rising a right hand, walking, waving a hand, running, sitting on the floor, lying down on the floor, jumping and getting down on the floor gestures. The model is based on the work of Rabiner [5]. It consists of two parts. One part is build of ergonic or fully-connected HMM. The ergonic part of the model has the task to extract the garbage movement. In this category all movements which are not to be detected are put in it. The second part of the model is a left-right model which detects the desired gestures (see figure 2). By training the model with help of the feature vectors, gesture can be extracted.
To test his system he took sequences of movements and compared it to substitution, deletion and insertion errors. Substitution error occurs when a gesture occurs and is detected instead of another. Deletion error occurs when a gesture is not even detected by the framework. Insertion error occurs when a movement is reported who did not occur. With help of these measurements he computed the reliability as follows: $reliability = \frac{correctlyrecognizedgesture}{deletionerror+insertionerror}$. He achieved a reliability of over 89% for each of the movements.

3.3 Classification with help of binary semantic classification tree

Lu [4] detects pointing, waving, raising a hand, describe width and describe height gestures with his method. He captures the gestures with a commercial motion capture system from Motion Analysis Corporation. Classification are done by a binary semantic classification tree (see figure 3).
His method enables him to use multiple classifier in putting each of them in a layer of the tree. In the first and third layer he puts a GentleBoost classifier and in the second layer he puts a k-nearest neighbor. By walking through the tree in top down order he proceeds in classifying. On the first layer he uses the velocity of elbow, wrist, hand and finger to separate left from wright hand gestures. In the second layer he uses the maximal velocity of a gesture trajectory to extract key posture. Since different persons have a slight different trajectory for the same gesture a K-means cluster cluster

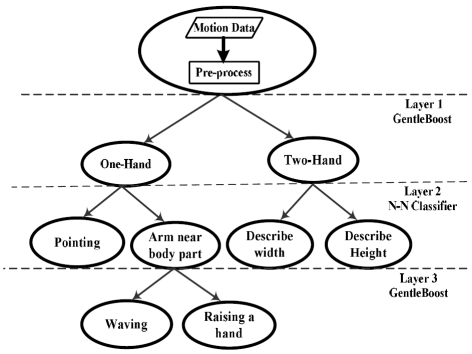

Figure 3: Classifiction by a binary semantic classification tree

this trajectory for each frame. These clusters are used as base for k-nearest neighbor algorithm to detect pointing, describe width and describe height gestures. In layer three, the detection is done by a GentleBoost algorithm over the periodicity of a gesture. With his help he separates waving and raising a hand gesture.
Experiments were done with 30 subjects. Each of them performed 5 categories of gesture in three times. 225 training sets and 225 testing sets where generated for his testing. A total result of 93.7 percent accuracy was achieved.

4. CONCLUSION

We saw in section two, methods able to localize changes in position of arm and body in relation to time. Section 2.1 used skin color detection to track gesture. Section 2.2 used velocity coupled with covariance differences to a scene to build a body representation. Section 2.3 uses an integration of different cue to track movement in 3D space.
It seems that the advantage of the method in section 2.1 to the method in section 2.2, is the shorter time to compute and the independence of the background. But method 2.1 is not able to represent a hole body since it detects only skin and shirt color. If we take the two method and compare them to the method of section 2.3 we see the advantage of the integration of different methods to form a detection method. First the method of section 2.3 is able to detect movement in 2D and translate them into 3D. The use of geometrical constraints, guarantees to find highly probable points where the arm is located. But as section 2.1 this method does not build a hole body.
In section three we saw methods to classify movements. Section 3.1 showed how significant informational gesture can be found. Section 3.2 and section 3.3 extracted predefined gestures. It is quite difficult to compare the methods of section 3, even if they have been measured with their failure rate. The difficulty of their comparison lies in their specific application. For example the gesture of section 3.2 and 3.3 are not the same. If one of this gestures is easier to detect than the detection rate, would give no information of the efficiency of the method.
If we pay attention at section 2 and 3 we can see that section 3 has a higher order of gesture recognition. The methods of section 3 do not only try to catch the location of a body part in time, but rather try to extract meaningful information out of it. We see this especially in section 3.2 where the coordinate (angles) of the body parts are used in a HMM to extract certain movement.
In general we saw the huge potential in recognition of arm and body gesture. It tends from human computer interface to behavioral pattern recognition to the simple the generation of an avatar. We see that the range of possible commercial application is quite extended.

5. REFERENCES

[1] Y. Azoz, L. Devi, and R. Sharma. Reliable tracking of human arm dynamics by multiple cue integration an constraint fusion. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 19(7):780–785, JULY 1997.

[2] N. Badler, M. Costa, L. Zhao, and D. Chi. To gesture or not to gesture: What is the question? *Computer Graphics International, 2000*, pages 3–9, 1992.

[3] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer. Technique for automatic emotion recognition by body gesture analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6. InfoMus Lab-Casa Paganini University of Genoa and Swiss Centre of Affective Sciences University of Geneva, June 2008.

[4] W. Lu, W. Li, L. Wang, and C. Pan. Gestures classification based on semantic classification tree. In *2nd International Congress on Image and Signal Processing*, pages 1–5. National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, October 2009.

[5] L. R. Rabiner. A tutorial on hidden markov models and selected application in speech recognition. *Proceedings of IEEE*, 77(2):257–286, February 1989.

[6] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, December 1992.

[7] S. Waldherr, R. Romero, and S. Thrun. A gesture based interface for human-robot interface. *Autonomous Robots*, pages 151–173, September 2000.

[8] A. D. Wilson, A. F. Bobick, and J. Cassell. Recovering the temporal structure of natural gesture. In *Proceedings of the Second International Conference on Automatic Face Gesture*. MIT Media Laboratory, October 1991.

[9] C. R. Wren, A. Azarbayejani, T. Darrel, and A. P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 19(7):780–785, JULY 1997.

[10] H.-D. Yang and S.-W. Lee. Reconstructing 3d human body pose from stereo image sequences using hierarchical human body model learning. In *The 18th International Conference on Pattern Recognition*. Department of Computer Science and Engineering, Korea University, 2006.

[11] H.-D. Yang, A.-Y. Park, and S.-W. Lee. Robust spotting of key gesture from whole body motion sequence. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*. Department of Computer Science and Engineering, Korea University, 2006.