Silhouette Extraction and Human Posture classification on still images

Michaël CODINA
Kent Ridge Digital Labs

Tele TAN
Kent Ridge Digital Labs

Philippe MULHEM
IPAL-CNRS

21, Heng Mui Keng Terrace
Singapore 119613
{codina,teletan,mulhem}@krdl.org.sg

Abstract

This paper discusses a novel approach for human body extraction and posture classification from still images. This step can be used to provide additional information about the image content, especially those with a strong human flavor. Our approach eliminates the need for image segmentation as the first step, thereby it does not suffer from the consequences of poor delineation of regions. We use a 2D puppet model to represent the human body form. A puppet is defined using a hierarchical and articulated graph. The optimization process is initiated by the input of the head, and then the position and characteristics of each adjacent limb is optimized using a local similarity measure. The set of attribute features of the limbs is used to describe the posture of the human subject. We tested the performance of these features with decision tree and artificial neural network classifiers. For a two class problem (i.e. posture of sitting and standing), we achieved a classification accuracy of 85% on a set of 320 realistic images.

1 Introduction

In the context of image indexing and retrieval engines, the presence of human is an important element in the representation of the images content. Because people describe images by considering objects and people [8], many queries involve the status or the relationships between a person and its environment in the image, as expressed in the query "I look for photographs with John waiting in front of the Eiffel Tower". This query, expressed in a natural language form, characterizes the configuration of a body by a class of posture (like "standing", "sitting", etc.). Such semantic symbols must be considered by image retrieval engines, because people describe images by symbols related to the actions or to the postures of the subject of an image.

Works in this field greatly use the temporal component of video sequences in a way to extract human silhouettes [3, 15]. Other approaches use the extracted silhouettes to reconstruct human bodies [12, 1].

When considering still images, few works intend to tackle the reconstruction of human silhouettes. David Marr in [7, p306], illustrates a pattern recognition approach on human silhouette recognition. The work of Yoshinari Kameda [6] use 3D model-matching of an articulated objects that can be used to recognize human bodies on synthetic images, but this kind of approach is very sensitive to noise in the processed images. Other approaches make use of special input devices, like thermal infrared cameras [4], or sets of cameras [5] to characterize and to track person gestures.

The work described in [10] intends to find the position of the limbs of a person. This approach was only demonstrated in simulated data, and according to our tests the skin color detector used in [10] is not accurate. This work is also based on a segmentation process without model, and uses geometrical forms and heuristics to build a simple human model.

Many research have been conducted on human gesture recognition (especially for interaction purposes) but, to the best of our knowledge, only [2] intends to qualify human postures in classes of postures (sitting, standing, lying and kneeling). This work determines additionally the point of view from which the human

is photographed (like "frontal" and "profile"). This approach achieves very good results (95% of correct description), but is based on a silhouette extraction process on video analysis.

Our concern here is to describe the class of posture of human subject on color photographic still images. Such approach needs first to extract the silhouette of the subject. This problem is very complex to achieve, but we show that the use of 2D puppets allows to find accurate approximation of the real silhouette of a subject. Puppets are defined using a hierarchical and articulated graph with constraints. The optimization process is initiated by the input of the head, and then the position of each limb is optimized using a local similarity measure. Additional hypotheseses consider that the whole body is visible in the image, and no external occlusion occurs. The reason that underlies the use of additional hypotheseses is that without such hypotheses it is currently almost impossible to go further the work of [9] where only standing pedestrian where able to be detected. However, our proposal is already robust against limited self occlusions. After the puppet of a human being is defined, we are able to use a decision step to characterize the class of the posture of the subject. The decision step is based on meaningful features from the extracted puppet. Currently, we are considering 2 classes of posture: sitting and standing. We show that the decision step using decision trees and feed-forward neural networks achieve high accuracy and are very stable.

The paper is organized as follows: in section 2, we present the puppet model and its optimization according to an input image, in section 3, we present the posture class decision step, in section 4, we apply our approach on a test set and we present the results obtained, in section 5, we conclude and we present the future directions of our work.

2 Puppet Extraction

In this section, we present the puppet extraction process. This process is based on searching for the optimal position of a 2D model of the puppet on the image. Before describing the puppet model proper, we would like to make the following remarks:

- The projection of a human body onto 2D image (noted I) destroys the information of depth and proportion. Also, we assume that it is difficult without other information (temporal, protected environment, etc.) to segment its silhouette accurately. To obtain, nevertheless, some information on its configuration, we thus propose an approximation of its silhouette by a 2D model (namely P) puppet, which is described in section 2.1.
- The configuration of the human body is obtained by a process of adjustment described in 2.3. This process optimizes the configuration of the puppet evolving in a 2D space (noted J) making possible to take into account the proportions of the human body. A stage of registration between I and J is described in section 2.2.

2.1 The model

The proposed model of the puppet, showed figure 1, is hierarchical, articulated, with constraints.

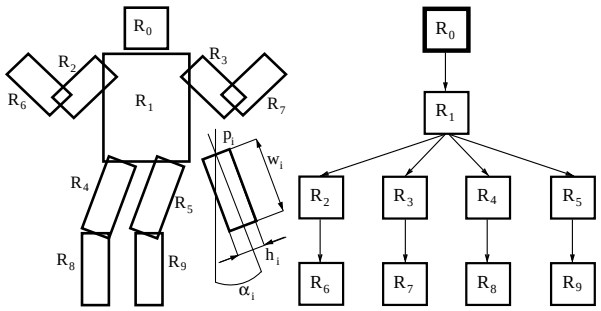

Figure 1. The hierarchical, articulated model of a puppet.

It is composed of 10 rectangles, each one localized in the plan J by a point of reference p_i, an angle with respect to the vertical α_i, a height h_i, a width w_i and an appearance function g_i. Therefore, the puppet can be represented as:

$$P = (R_0, \cdots, R_9) \text{ and } R_i = (p_i, \alpha_i, w_i, h_i, g_i) \; \forall i \in [0..9] \quad (1)$$

with $p_i \in \mathbb{R}^2$ and $\alpha_i, w_i, h_i \in \mathbb{R}$. g_i is a function with value in $\mathcal{A}$ which accounts the visual appearance of the members. $\mathcal{A}$ models color, texture, etc. Finally, a puppet is represented by 50 real parameters.

Because our model is articulated and to take into account the physical constraints of the human body, we fix the reference points $p_i \; \forall i \in [2..9]$ on their respective parents. These points represent the articulations of the puppet. The articulation of the trunk R_1 has nevertheless one degree of freedom: only x_{p_1} can move.

Because it is hierarchical, each member has a single relative. This relation of order in the elements of the model is presented figure 1. This hierarchy shows the significant role the trunk R_1 plays in the entire localization process. The primary members are the head

R_0 and the trunk R_1. The other secondary members are: arm $\{R_2, R_3\}$ and thighs $\{R_4, R_5\}$, followed by the front arm-lever $\{R_6, R_7\}$ and the legs $\{R_8, R_9\}$. Because it models a human entropological form, the angular and size characteristics of the puppet have constraints. They are modeled by a set of intervals within these constraints, i.e. $\forall i\ \alpha \in L_{\alpha_i},\ w \in L_{w_i}$ and $h \in L_{h_i}$. Because it is based on the appearance of the members of the human body, a set of 9 appearances functions $g_i : R_i \subset J \rightarrow \mathcal{A}$ which is used to primarly account for the clothing of a human subject.

2.2 Registration

Let us recall that we seek to position this 2D puppet onto the silhouette of a human subject. The first step is to anchor at least one of the puppet's member onto the corresponding point on the image. This position restricts the zone of study around this member of reference. Also, the scaling function can be derived from the registration process, i.e. $s : I \rightarrow J$. We thus make the assumption that all the characteristics of the head (R_0) are known. Localization of face from still images is sufficiently reliable today [14, 13] so that it is possible to estimate the localization of the head of a human subject under certain angular and size conditions.

2.3 Adjustment of the model

Within the framework of this paper, a local and hierarchical approximation of this articulated model has proved sufficiently effective to account for the real configuration of the human body. Taking into account the significant role of the trunk, the width w_1 and x_{p_1} were optimized. For the other members, w_i is fixed.

The number of characteristics entering in the process of optimization is thus finally 20: x_{p_1}, w_1 and $\alpha_i, h_i\ \forall i \in [1..9]$. Each characteristic is locally optimized in the following order: For R_1, x_{p_1}, α_1, w_1 then h_1 are optimized, then, $R_i\ \forall i \in [2..9]$ (the other members), α_i and h_i are optimized. To simplify our matter, we express this process within the framework of optimization of the angle α_i of the rectangle Ri.

The local optimization of a characteristic in this set is carried out by discretizing the values of the member's angle to some fixed step. n temporary rectangle $\mathcal{R}_{i_j}\ \forall j = [1..n]$ are thus calculated for α_j variable in L_{α_i}. We call $S = \bigcup_j \mathcal{R}_{i_j}$.

This optimization is based on the two values below. We define the precision, which measures how much the appearance of the interior of $\mathcal{R}_{i_j}$ is similar to what it should be:

$$p_p = \frac{|\{M \in S \mid (M \in \mathcal{R}_{i_j}) \wedge (\varphi(i, M) \leqslant p_{color})\}|}{|\{M \in S \mid \varphi(i, M) \leq p_{color}\}|} \tag{2}$$

(with $\varphi(i, M)$ a short notation for $\varphi(I(M), g_i(s(M)))$) and the specificity which measures, how much the exterior of $\mathcal{R}_{i_j}$ is dissimilar to what it should not be:

$$p_s = \frac{|\{M \in S \mid (M \notin \mathcal{R}_{i_j}) \wedge (\varphi(i, M) > p_{color})\}|}{|\{M \in S \mid \varphi(i, M) > p_{color}\}|} \tag{3}$$

An experimental parameter p_{color} is introduced here to take into account of the diversity of real images [1].

A measurement of similarity $\varphi : \mathcal{A}^2 \rightarrow \mathbb{R}$ is used. Because of the noise in the image and the inaccuracy of the model, the precision measure $p_p \ll 1$. In the same way, the specificity measure $p_s \neq 1$, because it is possible to find objects similar to the target outside of the zone. These two quantities evolve in the same way and account for the mapping of the rectangle $\mathcal{R}_{i_j}$ and the target. To obtain a single value, we multiply these 2 quantities:

$$p_{match} = p_p \times p_s \tag{4}$$

This formulation 4 remove any bias the specificity p_s would contribute. For example, when $p_s \approx 1$ and $p_p = 0$, the product is null for a zone which is obviously not similar. The process calculates all possible $\mathcal{R}_{i_j}$. We therefore select the highest p_{match} corresponding to a particular member characteristics.

3 Posture Decision

We seek to estimate the posture of human forms starting from still images. To test our approach we chose two most commonly used decision algorithms: probabilistic classifiers and neural networks.

We use the 18 following characteristics for the decision-making: the angle α_i and the quotient height over width h_i/w_i with $R_i\ \forall i \neq 0$.

3.1 Probabilistic Classifier

To test the decision of posture from probabilistic classifier, we use a decision tree generator. We first explore the C4.5 decision tree [11] implementation. This algorithm starts with large sets of cases belonging to known classes. The cases, described by any mixture of nominal and numeric properties, are scrutinized for patterns that allow the classes to be discriminated. It generates then a decision tree by minimizing the entropy of the set with to the method of Hunt. These

[1] Our experiments show that $p_{color} = 0.5$ gives good results.

patterns are expressed in the form of decision trees or sets of if-then rules, that can be used to classify new cases, with emphasis on making the model understandable as well as accurate. The decision trees generators use mainly logical or integer characteristics. The use of continuous values is made possible by discretisation.

3.2 Neural Network

We used a feed-forward neural network with 3 layers. The layer of input is consisted of the 18 numerical characteristics. At output, we await 2 binary values, for example "0 1" for "sitting", "1 0" for "standing". Several interpretations of the output are possible. We chose to threshold these 2 values against 0.5. The number of neurons of the hidden layer was fixed at 18.

4 Experimental results

4.1 Data Set

To test our approach, we obtained a set of 320 photographs that we used for training and testing purposes. To allow for a good varieties of postures and view points, we proposed the following data collection steps:

- 2 postures represented by sitting and standing.
- 2 camera heights (the photographer upright and squatted).
- 5 camera views of the subject (steps 45^o).
- 3 installations of the legs. In the case upright: "legs inserted, legs tightened, not stuck and legs perfectly stuck"; in the sitted case: "lengthened legs, legs folded at right angles and the bust leaned backwards.
- 4 installations of the arms: "isolated, along the body without contact, with contact and crossed (total proper occlusion)".

We thus obtain $2 \times 2 \times 5 \times 3 \times 4 = 240$ exhaustive photographs.

We then added 80 unspecified photographs while varying the distance to the camera, the posture, the position of the members and the environment.

4.2 Puppet extraction

With the process as described in section 2.3, we extracted the puppets from these 320 photographs on a PC Intel Pentium III 700Mh using "Windows Me" with 128Mb of RAM. For one image, this process took 1 minute for a study zone of 256×384 pixels. The localization of the head was previously performed manually.

Within the framework of this paper, we chose to approximate the appearance function g_i in the following manner: for the upper part ($R_i \ \forall i \in \{1,2,3,6,7\}$)), $g_i(M) = a_{upper} \ \forall M \in R_i$, and, for the lower part ($R_i \ \forall i \in \{4,5,8,9\}$), $g_i(M) = a_{lower} \ \forall M \in R_i$. These constants (a_{upper}, a_{lower}) were automatically approximated by an instance of the color in the middle of the trunk and under the belt.

We chose to use the color information based on HSV color model. Thus, φ is defined by $\varphi(a_1, a_2) = 1 - \|a_1 - a_2\|_2$, with a_1 and a_2 representing the couple hue (linearized and standardized) and the saturation (linearized) of the color.

The quality of the results of this extraction is not easily measurable in an objective manner, since it is not a real segmentation of the silhouette. However, from a qualitative point of view, we observed the results and decided in a subjective way to arrange these photographs in 3 classes according to whether the mapping were good, average or bad:

- The first class gives an account of a perfect matching between the silhouette and its puppet. The only noted errors may result from the inadequacy of the model.
- The second class is that of the cases where 1 or 2 secondary members (figure 1) deviates from the ideal position. This is often due to the similarity between the background and the subject.
- The last class accounts for the cases where the algorithm shows instability, mainly in the mapping of the trunk.

The table 1 shows the results obtained.

Good	Middle	Bad
68%	18%	14%

Table 1. Results of the classification

We obtain nearly 70% of good results. These good results are visually very close to the true silhouette and give an account of its configuration in 2D. The images of the middle class (18%) presented minor errors which do not seem to disturb the decision-making. The images of the bad class (14%) show the limits of the model and the algorithm of adjustment.

From a quantitative point of view, we compared the projection of the puppet in a binary image with the

optimal silhouette. These silhouettes were obtained beforehand by thresholding the 240 exhaustive photographs. For the 80 remaining photographs, we routed them by hand. This quantitative coefficient is the recovery rate between a silhouette S and a puppet P. The figure 2 shows two examples. This recovery rate is defined as follows:

$$r(P) = \frac{|(M \in P) \bigcap (M \in S)|}{|(M \in P) \bigcup (M \in S)|}$$

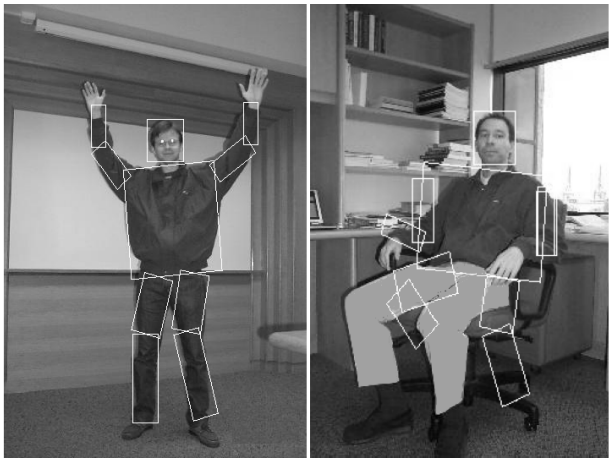

Figure 2. Examples of recovery rate: $r = 0.76$**(L),** $r = 0.43$**(R)**

We obtain an average rate of **62.8%** on the 320 images. This recovery rate can appear weak but it has to be compared with the best value that the model could perform. Indeed, given that w_i is fixed, $\alpha_i \in L_{\alpha_i}$, $h_i \in L_{h_i}$ and that the hands and feet are not modeled, r is much smaller than 1. We compare it to best puppets found whose recovery rate is roughly 80%.

4.3 Posture Decision

The decision of posture is easier to quantify. Having chosen 2 classes, any result must be compared with the mathematical chance of a random classifier of which the probability of giving the good answer is 0.5. We chose an experimental process called "leave one out" which consists of choosing one case c in n and to train the system on the $n-1$ remaining cases and to test on c. Results are showed in table 2:

We observe that the neural network is more accurate than the generator of decision tree. So, the neural networks are better at generalizing the decision model.

For C4.5, we measure the coherence of the training set by the calculation of a minimum error rate that the system generates when the input is perfect. This is done by launching the training and the test on the same set. The error rate obtained was 2.2%. For the neural network, the numerical error rate after training was 1.8%. It was stabilized after only 2000 iterations.

	Good Classification
C4.5	78.5%
Neural Network	**84.7%**

Table 2. Good Classification rates

4.4 Stability analysis

The results above show how our process of extraction of silhouette, coupled to a decision-making process is able to produce satisfactory global results with 84.7% of good classification.

We seek to measure now the intrinsic quality of the decision-making process. For that, the decision part should be launched on the best possible puppets. The error rate of the decision would be then a measurement of the quality of classification.

We chose to estimate this ratio by analysing the numerical limit of the recovery rate. For each extracted puppet P we calculate the decision error rate on the subset of the puppets whose recovery rate is lower than $r(P)$. Let us call $\mathcal{F}(x)$, the set of puppets for which $r(P) < x$. Among $\mathcal{F}(x)$ the classifier made mistakes on $\mathcal{F}_w(x) \subset \mathcal{F}(x)$, and the error rate is defined:

$$e(x) = \frac{|\mathcal{F}_w(x)|}{|\mathcal{F}(x)|} \quad (5)$$

The figure 3 shows the behavior of e.

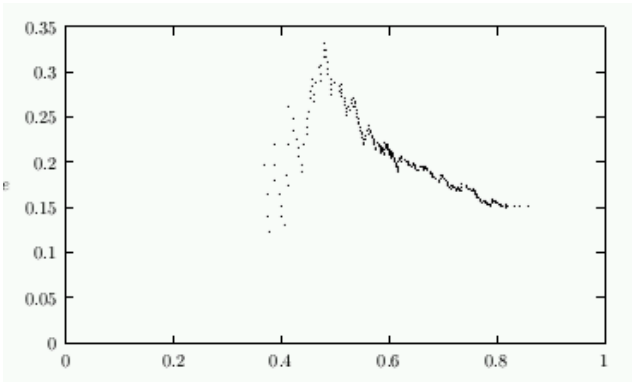

Figure 3. Behavior of e on the data set

First, the behavior of e for r less than approximately 0.48 is without any significance since the 45 cor-

responding puppets are not correctly extracted. The 45 puppets (14%) for which the recovery rate are less than 0.48 belong to the class "bad" as seen in section 4.2. Some of this error rate are below 0.15% which seems to be illogical. This behavior will be the subject of another study. For r greater than 0.5, e is decreasing, which shows the coherence of the decisional system used. In other words, the more the approximation of the silhouette is accurate, the more the decision is good: we had determined that the error rate converges to 0.15. Finally, the behavior of this error rate shows the relative stability of the decisional system. Indeed, improving r from 0.5 to 0.6 enhances the error rate by approximately 8%. Whereas improving r further 0.6 to 0.7 decreases the error rate by a lower margin of 3%. This implies that the decision system is robust to the errors of extraction of silhouette.

5 Conclusion & Future Works

The objective of the work described here is to extract an approximation of a silhouette from a human subject and to decide its posture from still images.

A 2D model, puppet, simulating the appearance of a human silhouette was introduced. A local and hierarchical approximation estimates the configuration of the human body. Based on this model, characteristics are provided to a probabilistic classifier and a neural network classifier. The best of the two classifiers used show a rate of correct classification of 84.7%.

The errors due to the approximate extraction of the silhouette are partly compensated by the decisional algorithm. The extraction of the silhouette, however estimated by a simple puppet appeared extremely satisfactory. The decision of posture, in the case of the neural network, is comparable to other works [2] which mainly starting from silhouettes extracted in video sequences.

Future works linked to this study are directed in 2 axis. First of all, it appears essential to us to evaluate the behavior of our system if the localization of the head is based on an automatic process of localization of faces. By construction, the adjustment process should be relatively resistant with possible imprecision. In a second time we want to generalize the model of puppet to take into account more realistic cases. In particular, the use of texture during measurement of the similarity should improve the precision of the extraction of puppet for subjects showing an unspecified clothing. Other generalizations will take into account minor external occlusions as much as other kind of posture like "lying" or "arms raised".

For the decision part, we think that a better extraction of the silhouette should bring more degrees of freedom in the input characteristics increasing the correct recognition.

References

[1] M. Brand. Shadow puppetry. In *ICCV99*, pages 1237–1244, 1999.

[2] I. Haritaoglu, D. Harwood, and L. S. Davis. Ghost: A human body part labeling system using silhouettes. *DARPA98*, pages 229–235, 1998.

[3] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Who, when, where, what: A real time system for detecting and tracking people. *Third International Conference on Automatic Face and Gesture, Nara*, pages 222–227, 1998.

[4] S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima. Real-time human posture estimation using monocular thermal images. *3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pages 492–487, 1998.

[5] S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, K. Ebihara, and S. Morishima. Human body postures from trinocular camera images. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 326–331, 2000.

[6] Y. Kameda, M. Minoh, and K. Ikeda. Three dimensional pose estimation of an articulated object from its silhouette image. *Proceedings of Asian Conference on Computer Vision*, pages 612–615, 1993.

[7] R. Marr. *Vision, A Computational Investigation into the Human Representation and Processing of Visual Information.* Freeman, 1982.

[8] C. Jörgensen. Attributes of images in describing tasks. *Information Processing and Management 34(2/3)*, pages 161–174, March/May 1998.

[9] C. Papageorgiou, T. Evgeniou, and T. Poggio. A trainable pedestrian detection system. *In Proceedings of Intelligent Vehicle*, pages 241–246, October 1998.

[10] J. S. Park, H. S. Oh, D. H. Chang, and E. T. Lee. Human posture recognition using curve segment for image retrieval. *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases 2000*, 3972:2–11, Jan. 2000.

[11] J. R. Quinlan. *C4.5: Programs for machine learning.* San Mateo:Morgan Kaufmann, 1993.

[12] R. Rómer and S. Sclaroff. Inferring body pose without tracking body parts. *IEEE Computer Vision and Pattern Recognition*, June 2000.

[13] H. Rowley, S. Balija, and T. Kanade. Neural network-based face detection. *IEEE PAMI*, 20(01):23–38, 1998.

[14] K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE PAMI*, 20(01):39–51, 1998.

[15] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.