Human Body Posture Recognition Using MPEG-7 Descriptors

Lutz Goldmann, Mustafa Karaman and Thomas Sikora

Technical University Berlin, Communications System Group,
Einsteinufer 17, Berlin, 10587 Germany

ABSTRACT

This paper presents a novel approach to human body posture recognition based on the MPEG-7 contour-based shape descriptor and the widely used projection histogram. A combination of them was used to recognize the main posture and the view of a human based on the binary object mask obtained by the segmentation process. The recognition is treated as a typical pattern recognition task and is carried out through a hierarchy of classifiers. Therefore various structures both hierachical and non-hierarchical, in combination with different classifiers, are compared to each other with respect to recognition performance and computational complexity. Based on this an optimal system design with recognition rates of 95.59% for the main posture, 77.84% for the view and 79.77% in combination is achieved.

Keywords: sensing people, posture recognition, action recognition, projection histogram, MPEG-7 contour based shape descriptor, minimum distance classifier, k-nearest neighbor classifier

1. INTRODUCTION

The "looking at people" research topic, that is, giving machines the ability to detect, track and identify people and their actions from video, has become a central topic in computer vision research. Examples include face recognition used for person identification, person detection, tracking and action recognition for surveillance applications, gesture recognition for smart interfaces and 3D tracking techniques for video coding and 3D image displays.

Our challenge is to develop a more general system using a unified hierarchical descriptor. The intention is to describe individual body parts and the person as a whole in order to deal with different analysis tasks, namely detection, tracking, action recognition and identification. The focus lies on a real-time system which operates on color imagery from a single camera. The MPEG-7 visual standard[1–3] offers a good framework for such an approach by providing suitable low level descriptors that allow us to measure similarity in images or image regions based on color, shape, texture or motion characteristics. The spatial arrangement of the different image regions or body parts characterized by various features is described by a special topology. Based on this, a typical analysis task can be seen as the extraction of a suitable feature vector in combination with a classification procedure. A further interpretation of the results on a higher level will lead to a semantic-related content description.

One important part of the overall system is the robust estimation of the human body posture. In combination with a temporal analysis it can be used for action recognition or event detection. Our approach for human body posture recognition is based on the MPEG-7 contour-based shape descriptor (CBSD), which appears to be a suitable descriptor for this shape analysis and classification task. In contrast to the region-based shape descriptor (RBSD), which is also part of the MPEG-7 standard, it describes only the closed contour of a single object instead of the entire region with holes or disjoint parts. Thus our approach is comparable with other silhouette-based approaches.[4, 5] Furthermore it is robust to significant non-rigid deformations and distortions along the contour due to perspective transformations, typical for persons appearing in image and video material.

In order to find an optimal approach for this pattern recognition task, various non-parametrical classification algorithms are examined. Beside the contour-based shape descriptor, the projection histogram is analyzed

Further author information: (Send correspondence to L.G.)
L.G.: E-mail: goldmann@nue.tu-berlin.de, Telephone: +49 30 314 25451

concerning its suitability for this task. Various non-hierarchical and hierarchical approaches are considered for this multi-class problem.

The paper is organized as follows. Section 2 reviews related work with special interest to human body posture recognition. Section 3 describes the overall system and the human body posture recognition part in detail. In section 4 the experimental results are presented. Section 5 summarizes and conlcudes our work.

2. RELATED WORK

The overall field of "looking at people" is quite broad. Gavrila[6] surveys the applications and recent developments in this domain. The scope is limited to the areas of action and gesture recognition. Thus the emphasis is on discussing the various methodologies used for the analysis of the human body and faces. Various approaches with different dimensions and models are discussed and appropriate systems are reviewed.

Much research work concentrates on human action recognition which corresponds to the analysis of human motion. Thereby spatial and temporal characteristics of an object need to be considered. The estimation of the human body posture and the localization of the body parts is one way to analyze the spatial part. The temporal part is mainly considered by analyzing specific features over time.

Ali[7] presents a system for automatic segmentation and recognition of continous human activity. It utilizes an explicit body model containing 3 body components (torso, upper leg, lower leg) and the formed angles to classify frames into breakpoint frames or non-breakpoint frames. The temporal analysis of these angles leads to a discrete action recognition.

A novel motion descriptor based on optical flow measurements and a suitable similarity measure for its matching is proposed by Efros.[8] In combination with a k-nearest neighbor classifier it is used to recognize human actions in various application scenarios. Beside the retrieval of simple action labels it is used to transfer skeletons onto the objects, synthesize data-based actions and clean action sequences from artifacts e.g. occlusions.

More closely related to the specific task of human body posture recognition is the the work of Nakajima.[9] He combines person identification with view recognition in a hierarchical system. Different low-level descriptors based on shape and color information are utilized, namely color histogram, projection histogram and local shape patterns. The recognition is carried out by using a hierarchy of bi-class support vector machines (SVM).

Fujiyoshi[5] proposes a modification of the widely used skeletons to a so-called "star" skeleton, which is applicable to humans and other objects as well. Human motion is broadly classified by determining the main posture using inclination of the skeleton and exploiting the temporal characteristics by cyclic motion analysis of its extremal points.

Ghost[4] uses a silhouette-based body model for localizing the main body parts (head, hands, torso and feet). First, horizontal and vertical projection histograms are used in combination with hierarchical minimum distance classifiers to estimate the main posture and the view point. Next, the localization of the body parts is accomplished by combining a convex hull analysis, partial mapping of the body parts and the known topology of the human body for a each main posture.

Existing approaches related to the field of human action recogntion can be broadly categorized into top-down[4, 7] and bottom-up[5, 9] approaches. Top-down approaches employ a explicit geometrical model of the human body to describe the human body posture. The model parameters are obtained by a fitting procedure, which yields rich high-level representations. Bottom-up approaches utilize visual low-level features which are classified into human action categories. In general, the description of actions with low-level representations is more difficult than with high-level representations, but their extraction process is more robust than model fitting procedures. In order to develop a more general and robust system, we utilize bottom-up methods based on low-level descriptors.

3. SYSTEM OVERVIEW

This section gives an overview of the overall system and describes in detail the part of the human body posture recognition. In figure 1 the basic structure of our "looking at people" system is outlined. Basically it consists of the following modules, which are described in the next sections: segmentation, detection, object description, classification and tracking.

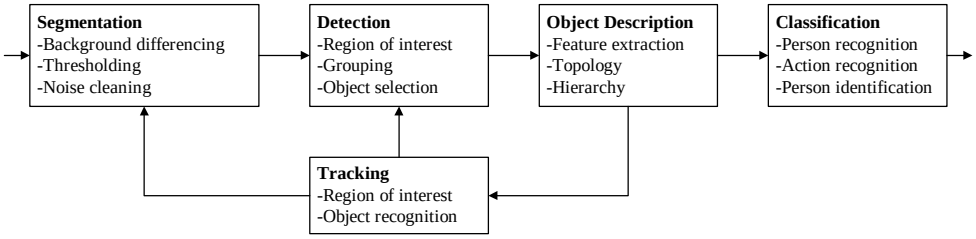


Figure 1. Overview of the overall system.

3.1. Object Segmentation and Detection

The segmentation and detection stages are often combined and simply called object detection. Basically the task of the segmentation is to split the image into several regions based on color, motion or texture information, whereas the detection stage has to choose relevant regions and assign objects for further processing. Figure 2 shows the main steps of our segmentation algorithm.

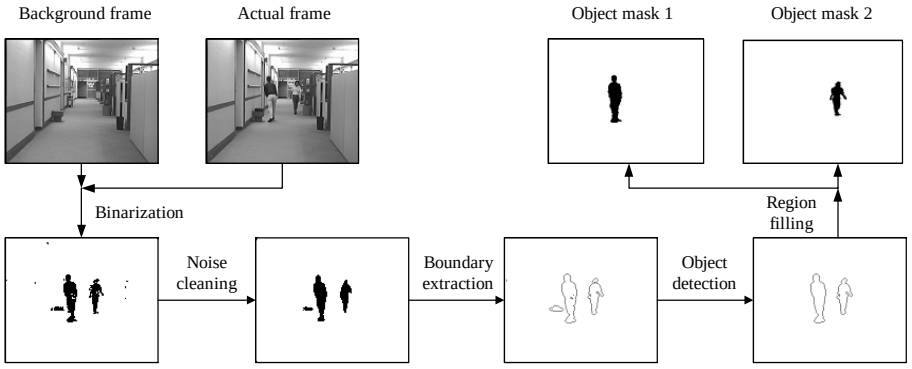


Figure 2. Overview of the segmentation and detection stage.

We follow the assumption that the video sequence is acquired using a stationary camera and there is only very little background clutter which is typical for an indoor environment. Based on this we use background differencing followed by thresholding to obtain a binary mask of the foreground region. In order to remove noise median filtering and morphological operations are used. Regions of interest (ROI) are detected using boundary extraction and a simple criteria based on the length of the boundary. Boundary filling is applied to each ROI and the resulting binary object masks are given to the description stage.

3.2. Object description

The object description refers to a set of features that describe the detected object in terms of color, shape, texture, motion, size etc. The goal of the feature extraction process is to reduce the existing information in the image into a manageable amount of relevant properties. This leads to a lower complexity and a more robust description. Additionally, the spatial arrangement of the objects within the video frame and as related to each other is characterized by a topology.

A very important issue for the performance of a subsequent classification task is to select a suitable descriptor that expresses both the similarity within a class and the distinctions between different classes. Since the classifier strongly depends on the information provided by the descriptor it is necessary to know its properties and limitations relating to this specific task. In case of human body posture recognition it is obvious that shape descriptors are needed to extract useful information. This section outlines two descriptors that are used throughout our erxperiments.

3.2.1. Projection histogram

A widely used descriptor[4,9] for shape analysis tasks is the projection histogram (PH). It describes the region of an object by the use of pixel projections onto the cartesian coordinate axes. The normalization is done by rescaling the object mask to a maximum size of 100 pixels in one dimension and centering a 100x100 pixels window on it. The vertical and horizontal histograms are obtained from the object mask by counting the pixels along the horizontal and the vertical lines respectively.

In our system the descriptor consists of two histograms with 100 bins each. That leads to a feature vector length of 200. The projection histogram is scale invariant, but sensitive to rotation or mirroring. To compare two projection histograms with each other the different Minkowski metrics (max, Manhattan and Euclidian distance) can be used.

3.2.2. Contour-based shape descriptor

The contour-based shape descriptor (CBSD) is one of the shape descriptors which are part of the MPEG-7 visual standard.[1–3]

It is based on the curvature scale-space (CSS) representation of an object contour. The CSS representation decomposes the contour into convex and concave sections by determining the inflection points (e.g. points at which the curvature is 0). This is done in a multiresolution fashion, where the contour is analyzed at various scales, obtained by a smoothing process. This can be illustrated by the so called CSS image, which shows how the inflection points (peaks) change during the iterative filter operation. The CSS x-axis corresponds to the position along the contour (clockwise, starting from an arbitrary point) and the y-axis corresponds to the magnitude which is the amount of smoothing iterations needed to remove the peak.

The descriptor consists of an index indicating the number of peaks (maximum 64) in the CSS image, the global and prototype curvature vectors that are built by the eccentricity and circularity values of the original and filtered contour, the magnitude of the largest peak and the positions and magnitudes of the remaining peaks. This yields a maximum feature vector length of 134. The CBSD has some very useful properties.[10] Firstly, the representation is invariant to rotation, scaling and mirroring the object contour. It has also been shown that it is robust to noise on the contour and in case of rigid, non-rigid or perspective deformations. Finally, the CBSD can efficiently differentiate between shapes that have distinguishing properties in their contour, while having similar region properties. Two CBSD's can be compared to each other by the use of a special distance metric, which is recommended as non-normative part in the MPEG-7 standard. Due to the involved structure of the feature vector the use of simple distance metrics is not useful.

3.3. Classification

Classification is a pattern recognition (PR) problem of assigning an object to a class. Thus the output of the PR system is an integer label. The task of the classifier is to partition the feature space into class-labeled decision regions. Basically, classifiers can be divided into parametric and non-parametric systems depending on whether they use statistical knowledge of the observation and the corresponding class. A typical parametric system is the gaussian mixture modell (GMM), which assumes Gaussian distribution of each feature in the feature vector. An example for non-parametrical systems is the so-called k-nearest neighbor (k-NN) classifier. In the following the two classifiers used in our experiments are described.

3.3.1. Minimum distance classifier

The minimum distance classifier (MDC) belongs to the class of non-parametric pattern recognition algorithms. It's also known as template matching, since a template for all the training samples of one class is built by element-wise mean calculation of the feature vectors. Given an unlabeled test sample it searches for the closest class template and assigns its corresponding class.

Possible implementation options include the used distance metric and the way the templates are calculated. The main advantages are the marginal memory requirements (for each class only one feature vector is stored) and the low complexity of the classification step (the number of distance calculations is equal to the number of classes). Disadvantages include the slightly increased complexity during the training step and the often inadequate class model which leads to bad recognition results.

3.3.2. k-nearest neighbor

The k-nearest neighbor classifier (k-NN) belongs also to the non-parametric pattern recognition methods. It is a very intuitive method that classifies unlabeled test samples based on their similarity to labeled training samples. For a given unlabeled sample it finds the k closest samples in the training data set and assigns the class label that appears most frequently within the k-subset. If multiple classes appear with the same frequency the label of the class with the smallest distance is assigned.

Parameters to adapt include the number of nearest neighbors k, the distance metric and the used search procedure. The advantages of the k-NN classifier are its simple structure which makes it easy to analyse, and the low complexity during the training step. The main disadvantages of k-NN classifiers are the huge memory requirements (each training sample is stored with the corresponding class label) and the complexity of the classification process (the number of distance calculations is equal to the number of training samples). Various methods (bucketing, kd-tree, bd-tree) exist for speeding up the search within the classification step. We consider only the simple brute force approach for our purpose.

3.4. Human Body Posture Recognition

Based on our overall system approach we treat the human body posture recognition as a basic classification task. Given a novel binary object mask to be classified, and a database of samples labeled with possible body postures, the previously described shape descriptors are extracted and the image is classified with a choosen classifier. This can be interpreted as a database query: given a query image, extract suitable descriptors and retrieve the best matching human body posture label. Other similar queries are possible, resulting in a number of useful applications, such as skeleton transfer, body posture synthesis and figure correction. In the following, human body posture classes are defined and the structure of the classifier is outlined.

3.4.1. Main postures

People can be in many different postures while they are performing actions. Primarily the postures can be classified into four main postures,[4] namely standing, bending, sitting, laying. The discontinous transition between two classes is set intuitively in the middle of the continous transition between representative postures of each class. Figure 3a shows typical examples for the different main postures and the defined transitions.

3.4.2. Views

Furthermore each posture has different appearances (views) varying with the view point of the camera. These views are typically divided[9] into front-view (0°), right-view (90°), back-view (180°) and left-view (270°). As for the main postures we set the transitions intuitively at angles of $(45°, 135°, 225°, 315°)$. Figure 3b shows typical examples for the different views and the defined transitions.

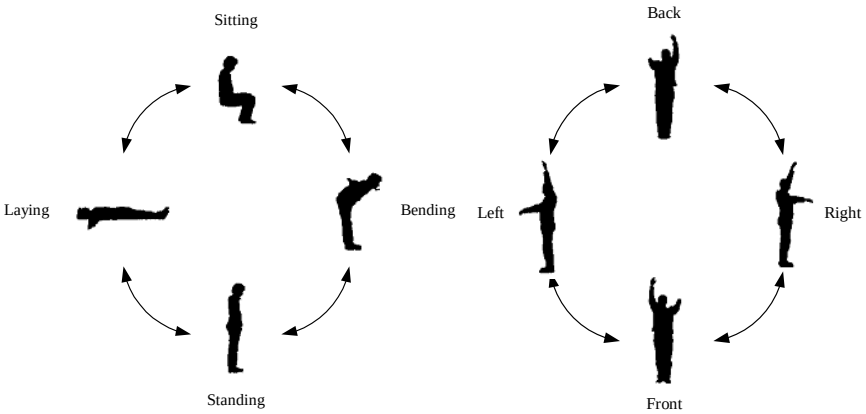
3.4.3. Advanced postures

Beside the main postures several advanced postures can be defined that emerge from the independent movement of the primary (head, hands, feet, and torso) and secondary body parts (elbows, knees, armpits, hip, upper back).[4] This yields a large number of variations, which are difficult to map into classes. An alternative approach is to localize the body parts and match them to a model of the human body. This will be task for future exploration.

3.4.4. Recognition process

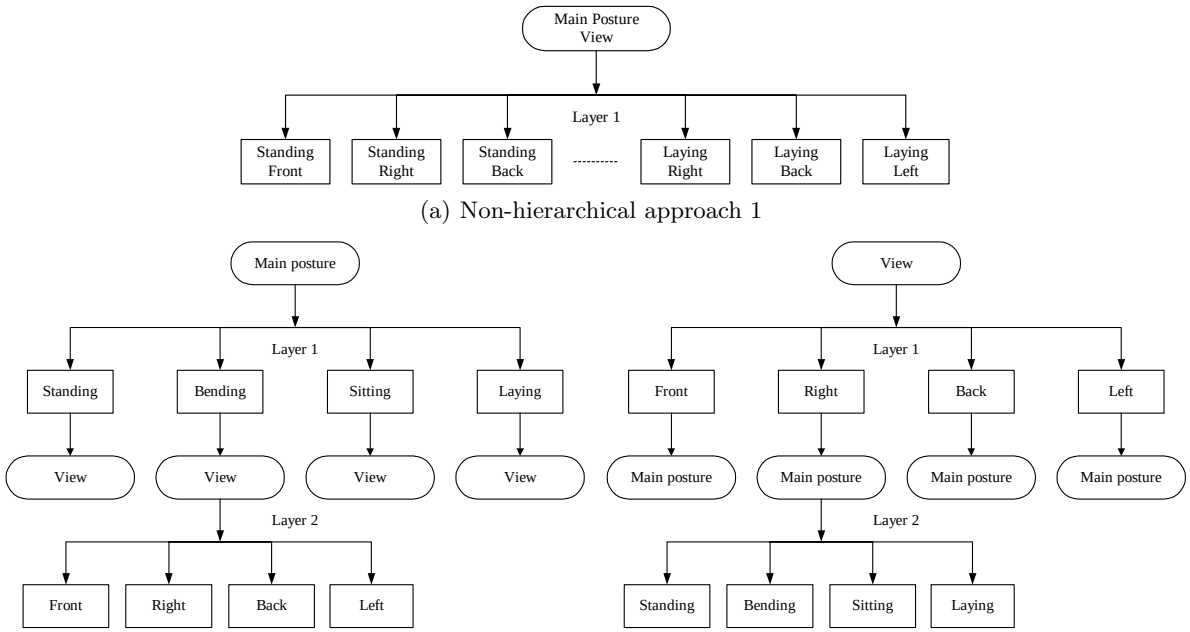
The actual goal is to assign a main posture and a view to a given test sample. We can treat this as one classification task in a non-hierarchical manner or use hierarchized classifiers. The latter case allows two structures that differ in the order of the individual classification tasks. All approaches are shown in figure 4.

The non-hierarchical approach 1 needs a single 16-class classifier which assigns a main posture and a view at the same time. Both hierarchical appproaches require 5 4-class classifiers and thus allow two combinations. In approach 2 the first classifier assigns a main posture and then the corresponding second classifier assigns a view. Approach 3 is just the inverse of approach 2. Table 1 gives an overview of the assigned class labels and the indices used in the figures throughout the paper. The index 0 corresponds to an undefined class.



(a) Main postures. (b) Views.

Figure 3. Partition of the human body postures and their transitions.



(a) Non-hierarchical approach 1

(b) Hierarchical approach 2 (c) Hierarchical approach 3

Figure 4. Different classification approaches for human body posture recognition.

Table 1. Overview of the class labels and its indices.

Main posture \ View		Front	Right	Back	Left
		1	2	3	4
Standing	1	1	2	3	4
Bending	2	5	6	7	8
Sitting	3	9	10	11	12
Laying	4	13	14	15	16

Table 2. Recognition rates in % for different validation procedures used to evaluate the main posture classification.

Validation methods	MDC	k-NN			
		k=1,2	k=3	k=4	k=5
Resubstitution	73.64	100.00	97.86	97.86	96.89
Holdout method	74.35	94.16	92.53	93.18	90.90
k-fold cross validation	73.82	95.59	94.90	95.39	94.71

4. EXPERIMENTS

This section summarizes the experiments, that have been conducted to evaluate the approaches mentioned in section 3.4. An extensive database of video sequences containing 4 persons, 4 main postures, 4 views and all possible advanced postures was created. The video sequences were captured with a stationary color camera at a resolution of 352x288 pixels (CIF Format) and a framerate of 15 fps. The object masks were obtained by the segmentation and detection stage and hand-labeled by a human observer in order to set the ground truth.

Based on this the classification performance of the different implementations, is measured using the recognition rate (RR) which is defined as the ratio of correct classified samples to the total number of test samples. For further analysis several other measures and representations of the classification results are used.

The classification matrix[8] shows the relationship between the labeled (ground truth) data and the classified data for all available classes. Each row represents the probabilities of a certain class to be classified into these classes. The class index 0 is used to indicate an undefined class if the classifier is not able to assign a valid class label to the test sample. The main diagonal shows the propabilities of correct classified samples for each class which corresponds to the true positive (TP) rate. Analysis of the classification matrix can reduce the confusion among classes.

The computational complexity is measured as the time which is needed for one classification procedure and alternatively for one distance calculation. It is measured on an AMD Athlon processor with 1.47 GHz.

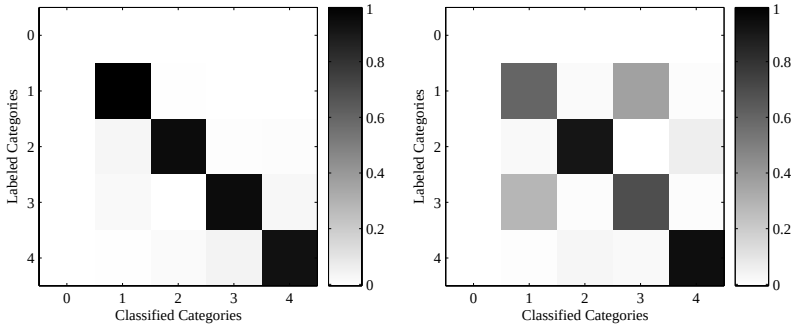
Various methods exist for the validation of classification methods. The most prominent are resubstitution (also called testing on training data), the holdout method and the k-fold cross validation. We conducted some simple experiments to analyze the characteristics of the different validation methods. The resubstitution method delivers often very unrealistic results since the testing data is also used for the training of the classifier. This yields recognition rates which are better than the realistic ones. This is especially true for a 1-nearest neighbor classifier, because the validation using resubstitution gives always a recognition rate of 100%. The holdout method overcomes this problem by splitting the data set randomly into test and training data. Thus the results strongly depend on the actual splitting of the data set. For further experiments we use the k-fold cross validation method because it gives the most realistic results due to the fact that each sample of the data set is used independently for training and testing. This is confirmed by table 2 where the results of the k-fold cross validation lie between the results of the resubstitution and the holdout method.

The next experiments were conducted to understand what the different descriptors and classifiers can achieve regarding the individual classification tasks. In table 3 the classification results for the projection histogram with different distance metrics are given. As we can see the 1-nearest neighbor classifier performs better than the other classifiers in both tasks. While it achieves a good recognition rate of RR = 95.59% for the main posture it reaches only a fair recognition rate of RR = 78.53% for the view. Especially the front and back view is confused quite often as depicted in figure 5. The distance calculation times are 0.021 ms (max distance), 0.017 ms (Manhattan distance) and 0.060 ms (Euclidian distance), respectively. Because it achieves the best recognition rate with in the lowest calculation time the Manhattan distance is found to be the optimal distance metric for our tasks. The classification of one test sample with a 1-NN classifier and 1000 training samples takes about 0.017s.

We performed a similar experiment for the contour-based shape descriptor and the results are shown in table 4. Again the 1-nearest neighbor classifier yields the best results for both tasks and the performance for the main posture with RR = 90.39 is superior to that of the view RR = 60.00%. As expected the distance metric defined

Table 3. Recognition rates in % for main posture and view classification tasks using the projection histogram.

Task	Distance Metric	MDC	k-NN			
			k=1,2	k=3	k=4	k=5
Main posture	Max	53.72	82.16	80.29	80.39	77.94
	Manhattan	73.82	95.59	94.90	95.39	94.71
	Euclidian	73.82	94.90	93.63	93.63	92.65
View	Max	37.55	61.96	57.75	57.45	55.20
	Manhattan	37.84	78.53	74.22	76.08	73.63
	Euclidian	40.39	74.41	70.00	71.67	68.82



(a) Main postures (95.59%). (b) Views (78.53%).

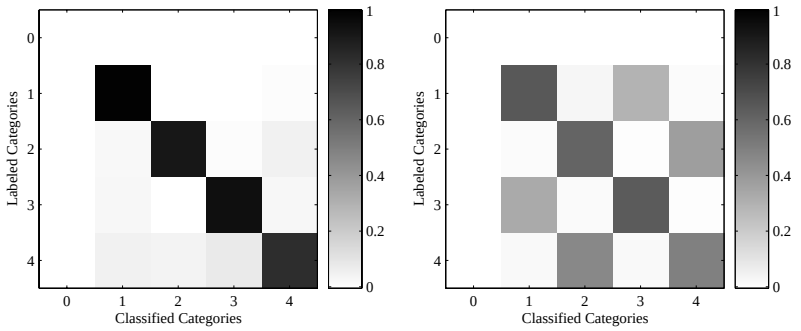
Figure 5. Classification matrices for main posture and view classification task using the projection histogram.

in the MPEG-7 standard causes a strong confusion between opposite views (see figure 6), due to the mirroring invariance. To solve this problem the distance metric has been modified in order to make it sensible to mirroring. This causes a small performance loss of 3% for the main posture but we obtain a significant gain of 18% for the view. Based on these results we use the MPEG-7 distance for classifying the main posture and the modified distance for classifying the view. The distances also differ in its calculation time. While calculating the MPEG-7 distance takes about 0.893 ms calculating the modified distance takes only 0.394 ms. Thus one classification task with a 1-NN classifier and 1000 training samples takes about 0.9 s and 0.4 s, respectively. The invalid results for the minimum distance classifier are caused by the special structure of the CBSD feature vector in combination with the class templates generated by the MDC. If such a template is matched (calculate the distance) against a test sample this leads to an invalid distance value.

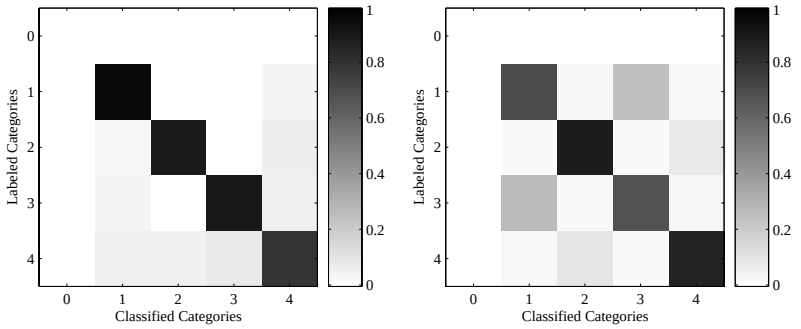
Based on this preliminary experiments the evaluation of the approaches mentioned in section 3.4 has been accomplished. In case of the contour-based shape descriptor the MPEG-7 defined distance metric has been used for the main posture and the modified distance metric for the view. For the projection histogram the Manhattan distance has been utilized for both tasks. The hierarchical approaches make it possible to use different descriptors on the individual layers. If only one descriptor is used we call this a single descriptor classification and otherwise

Table 4. Recognition rates in % for main posture and view classification tasks using the contour-based shape descriptor.

Task	Distance metric	MDC	k-NN			
			k=1,2	k=3	k=4	k=5
Main posture	MPEG-7	0.00	90.39	89.02	89.02	89.90
	Modified	0.00	87.75	86.47	86.56	85.20
View	MPEG-7	0.00	60.00	58.43	58.33	57.65
	Modified	0.00	77.84	75.39	76.37	72.65



(a) Main pos., MPEG-7 (90.39%). (b) Views, MPEG-7 (60.00%).



(c) Main pos., Mod. (87.75%). (d) Views, Mod. (77.84%).

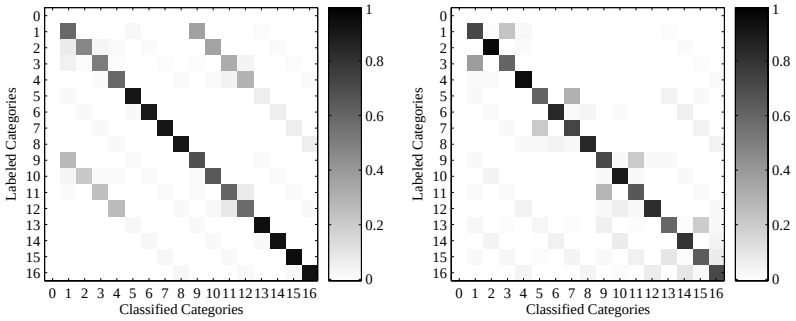
Figure 6. Classification matrices for main posture and view classification task using the contour-based shape descriptor.

Table 5. Recognition rates in % for main posture and view classification tasks using different approaches and combinations of descriptors.

Combination	Descriptors	Non-hierarchical	Hierarchical	
		Approach 1	Approach 2	Approach 3
Single	PH	75.98	76.80	77.27
	CBSD	72.07	75.47	71.86
Multiple	PH-CBSD	Not applicable	79.77	72.88
	CBSD-PH	Not applicable	72.21	76.01

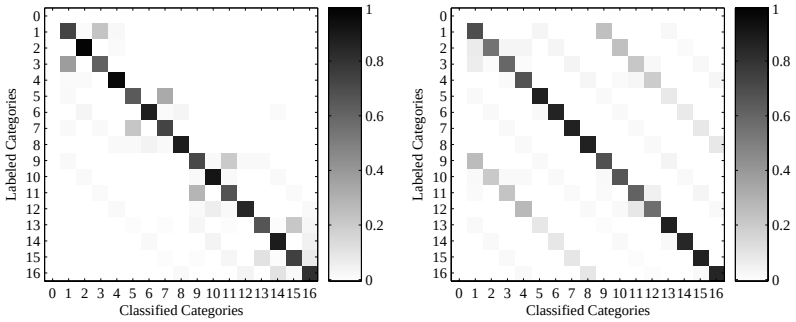
multiple descriptor classification. The evaluation results are summarized in table 5.

The hierarchical approaches outperform the non-hierarchical ones. If a single descriptor is used throughout the hierarchy, approach 3 gives the best results (RR = 77.27%) for the projection histogram, whereas approach 2 gives the best results (RR = 75.47%) for the contour-based shape descriptor. Figure 7 depicts the corresponding confusion matrices. Using multiple descriptors the recognition rates can be further impoved. Approach 2 using PH on layer 1 and CBSD on layer 2 gives the best results (RR = 79.77%) followed by approach 3 using CBSD on layer 1 and PH on layer 2 (RR = 76.01%). These results show that the projection histogram is more appropriate for classifying the main posture whereas the CBSD is more suitable for classifying the view. Analyzing the confusion matrices in figure 8 and 7 provides several interesting aspects. As noticed before the most confusion among the classes arises between the front and the back view, more than between the left and the right view. This coincides with the human recognition which utilizes known constraints (statistics) of the observed object. Suppose one sees the shadow of a human from a side view and one arm is raised to the front, it is easy to distinguish between left and right view. On the other hand if one sees a human from the front or back view and one arm is raised to the side, it is still difficult to tell which view it is. Another aspect is the confusion between



(a) App. 3, PH (77.27%). (b) App. 2, CBSD (75.47%).

Figure 7. Classification matrices for single descriptor classification.



(a) App. 2, PH-CBSD (79.77%). (b) App. 3, CBSD-PH (76.01%).

Figure 8. Classification matrices for multiple descriptor classification.

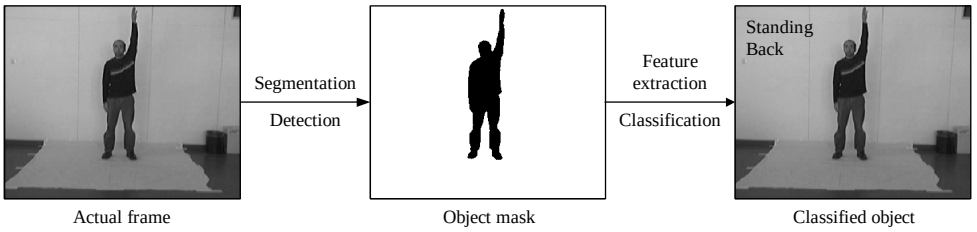
standing and sitting which is more obvious for approach 3. Possibly this is caused by the similar contour in front and back views. Especially the distinct contour of the head and the shoulders is present in both views.

Finally, in figure 9 some examples of our human body posture recognition system are shown. Example 1 shows a typical sample with confusion between the front and the back view. But if only the object mask is considered, even for a human it would be difficult to decide between these two views. Example 2 depicts a sample were both the main posture and the view are properly classified.

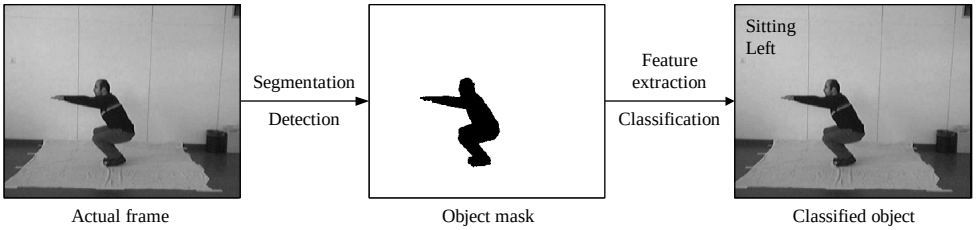
5. CONCLUSION

In this paper we presented a new approach to human body posture recognition. It is based on the MPEG-7 shape descriptor and the projection histogram. A combination of them was used to recognize the main posture and the view of a human based on the binary object mask obtained by the segmentation process. For this multi-class problem different non-hierarchical and hierarchical approaches were examined and the system structure was outlined. Although the results in recognizing the main posture are promising (RR = 95.59%), the recognition of the views remains a problem (RR = 77.84%). In combination with a hierarchical classifier this leads to a recognition rate of 79.77%. Possible improvements can arise using color or texture information to classify the view. This, and the recognition of advanced postures, which complies to localizing the body parts is part of future work.

The human body posture recognition is part of a "looking at people" system, which is currently under development. By combining the human body postures with temporal characteristics action recognition becomes possible. We are going to extend our main idea of using visual low-level descriptors in combination with suitable classification algorithms by applying it to other analysis tasks such as person identification and action recognition. Thereby we will concentrate on suitable descriptors and more sophisticated classification methods as well.



(a) Example 1.



(b) Example 2.

Figure 9. Some classification examples of our human body posture recognition system.

ACKNOWLEDGEMENTS

This work was supported by the DFG as part of the research initiative SPP Nr. 1041 "Distributed Processing an Delivery of Digital Documents".

REFERENCES

1. A. Yamada, M. Pickering, S. Jeannin, L. Cieplinski, J. R. Ohm, and M. Kim, "MPEG-7 Visual part of experimentation model version 10.0," Tech. Rep. N4063, ISO/IEC JTC1/SC29/WG11, 2001.
2. L. Cieplinski, W. Kim, J.-R. Ohm, M. Pickering, and A. Yamada, "MPEG-7 Visual part of experimentation model version 11.1," Tech. Rep. M7691, ISO/IEC JTC1/SC29/WG11, 2001.
3. L. Cieplinski, M. Kim, J. R. Ohm, M. Pickering, and A. Yamada, "Information technology – Multimedia content description interface – Part 3: Visual," Tech. Rep. N4062, ISO/IEC JTC1/SC29/WG11, 2001.
4. I. Haritaoglu, D. Harwood, and L. S. Davis, "Ghost: A human body part labeling system using silhouettes," in *Proc. of the 14th Intl. Conf. on Pattern Recognition (ICPR98)*, **1**, pp. 77–82, 1998.
5. H. Fujiyoshi and A. Lipton, "Real-time human motion analysis by image skeletonization," in *Proc. of the Workshop on Application of Computer Vision*, Oct 1998.
6. D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding* **73**, pp. 82–98, Jan 1999.
7. A. Ali and J. K. Aggarwal, "Segmentation and recognition of continuous human activity," in *IEEE Workshop on Detection and Recognition of Events in Video*, pp. 28–, IEEE, 2001.
8. A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. of the Intl. Conf. on Computer Vision (ICCV03)*, IEEE, 2003.
9. C. Nakajima, M. Pontil, and T. Poggio, "People recognition and pose estimation in image sequences," in *Proc. of IEEE-INNS-ENNS Intl. Joint Conf. on Neural Networks*, 2000.
10. B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7*, John Wiley & Sons, 1 ed., 2002.
11. J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding* **73**(3), pp. 428–440, 1999.
12. A. M. Baumberg and D. C. Hogg, "An efficient method for contour tracking using active shape models," in *Proc. of the Workshop on Motion of Non-rigid and Articulated Objects*, pp. 194–199, IEEE Computer Society, April 1994.

13. C. Cedras and M. Shah, "Motion-based recognition: A survey," *Image and Vision Computing* **13**, pp. 129–155, Mar 1995.
14. Y. Ivanov, C. Stauffer, A. Bobick, and W. E. L. Grimson, "Video surveillance of interactions," in *Proc. of 2nd Intl. Workshop on Visual Surveillance*, 1999.
15. M. Leung and Y. Yang, "A human body outline labeling system," *Trans. on Pattern Analysis and Machine Intelligence* **17**(4), pp. 359–377, 1995.
16. T. Mori, K. Tsujioka, M. Shimosaka, and S. T., "Human-like action recognition system using features extracted by human," in *Proc. of the Intl. Conf. on Intelligent Robots and Systems (IROS02)*, 2002.
17. A. Pentland, "Looking at people: Sensing for ubiquitous and wearable computing," *Trans. on Pattern Analysis and Machine Intelligence* **22**, pp. 107–119, Jan 2000.
18. A. Selinger and L. Wixson, "Classifying moving objects as rigid or non-rigid without correspondences," in *Proc. of the DARPA Image Understanding Workshop*, DARPA, 1998.
19. P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. of the Intl. Conf. on Computer Vision ICCV03*, Oct 2003.
20. C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *Trans. on Pattern Analysis and Machine Intelligence* **19**, pp. 780–785, Jul 1997.