NEW IMPROVED FEATURE EXTRACTION METHODS FOR REAL-TIME HIGH PERFORMANCE IMAGE SEQUENCE RECOGNITION

Gerhard Rigoll, Andreas Kosmala

Gerhard-Mercator-University Duisburg
Faculty of Electrical Engineering
Dept. of Computer Science
D-47057 Duisburg, Germany
{rigoll,kosmala}@fb9-ti.uni-duisburg.de

ABSTRACT

This paper describes new feature extraction methods which can be used very effectively in combination with statistical methods for image sequence recognition. Although these feature extraction methods can be used for a wide variety of image sequence processing applications, the target application presented in this paper is gesture recognition. The novel feature extraction methods have been integrated into an HMM-based gesture recognition system and led to substantial improvements for this system. It turned out that the new features are not only able to describe the gesture characteristics much better than the old features, but additionally they also led to a dramatic reduction in dimensionality of the feature vector used for representing each frame of the image sequence. This resulted in the fact that it was possible to use the novel features in combination with a new architecture for statistical image sequence recognition. The result of this investigation is a high performance gesture recognition system with significantly improved recognition rates and real-time capabilities.

1. INTRODUCTION

Image sequence recognition is a difficult subject with a large variety of possible applications. One of the most popular applications is gesture recognition, e.g. recognition of characteristic movements as hand waving, spin, pointing, or head movements. Gesture recognition is considered as one of the most promising areas of human-computer-interaction, e.g. for communication with multi-media systems or interaction with virtual environments. There are various approaches to this complex problem. The approach presented by the authors in [1,2] makes use of statistical methods for image sequence recognition. In this approach, each frame of the image sequence is decomposed into vertical or horizontal stripes. Each stripe is represented by the grey values of the pixels contained in this segment of the image. These grey values are summarized in a vector, which is transformed into a discrete label by a vector quantizer (VQ). In this way, each single image is represented by a sequence of VQ labels and the entire image sequence is represented by a corresponding sequence of VQ labels with the length equal to the number of the frames in the image sequence times the number of stripes in each single image. Several video sequences from different persons performing the same gesture are then recorded for each gesture and used as training data for the gesture recognition system. For each gesture, a discrete hidden Markov model (HMM) is trained with the VQ label streams representing the training data for this gesture. For recognition, the HMM's are used in order to calculate the most probable gesture with the Viterbi algorithm. Optionally, it is also possible to use probabilistic neural networks (see [1]) for training and recognition based on the VQ label streams.

The major outcome of the investigation of the original system in [1] was the fact that statistical pattern recognition methods are very suitable for image sequence recognition and that a system based on such an approach works amazingly well. It was therefore decided to further concentrate on the statistical approach for image sequence recognition and to further improve the system described in [1]. The result of these activities was a completely redesigned system for image sequence recognition with several major improvements which can be summarized in the following 4 points:

- New feature extraction methods more suitable for describing the gesture characteristics and dynamic aspects of the moving parts of the image and leading to a dramatic reduction in feature vector dimension.
- Concentration on the use of hidden Markov models rather than neural networks for dynamic pattern processing
- Switch from discrete to continuous HMM's, which has been made possible through the new feature extraction and the resulting reduction in feature dimension.
- Concentration on 2 different person-independent gesture recognition taks: The first task consists of 12 different gestures (see left part of Table 1) and the 2nd task containing 12 additional more confusable gestures (see right part of Table 1), resulting in a very complex and demanding gesture recognition task.

	Gestures for Task 1		Additional Gestures for Task 2
1.	HAND-WAVING-BOTH	13.	HAND-WAVING-RIGHT
2.	TO-RIGHT	14.	HAND-WAVING-LEFT
3.	TO-LEFT	15.	TO-BOTTOM
4.	TO-TOP	16.	STOP
5.	ROUND-CLOCKWISE	17.	NOD-YES
6.	ROUND-COUNTERCLOCKWISE	18.	CLAPPING
7.	COME	19.	TURN-RIGHT
8.	NOD-NO	20.	TURN-LEFT
9.	KOTOW	21.	DRAW-A
10.	SPIN	22.	DRAW-B
11.	GO-LEFT	23.	DRAW-C
12.	GO-RIGHT	24.	DRAW-D

Table 1: Gestures used in the improved gesture recognition system

The result from these improvements is a high performance real-time gesture recognition system with 99.5% recognition rate for Task 1 and 91,7% for Task 2, as described before. The system has not only achieved these very high recognition rates in person-independent mode for off-line recognition using a pre-recorded database, but has also proved to be surprisingly robust in real-world tests, while keeping its real-time capabilities. It has been recently demonstrated as research prototype at various exhibitions, including the world's largest industrial fair in Hannover, and has worked very reliably even with unexperienced users at the fair. Considering the robustness, the high recognition rates and the task complexity with up to 24 different gestures, we believe that this is now one of the most advanced gesture recognition systems which is based on pure image processing techniques, not assisted by any further tools, such as e.g. data gloves or colour markers.

This capability is the result from mainly 2 facts: The 1st fact is the use of statistical methods for dynamic pattern recognition problems, which has been already shown in [1] to be a very useful approach to gesture recognition. The 2nd fact is clearly the improved feature selection which resulted in a substantial improvement of the new system compared to the original system presented in [1]. This improved feature extraction method will be now further discussed.

2. NEW FEATURE EXTRACTION METHOD

One of the main drawbacks of the original approach was the relatively simple feature extraction method consisting of the transformation of the image sequence into a VQ label stream as described above. This method was effective but had 2 major problems: Firstly, such a feature extraction did certainly not emphasize any characteristics of the actual gesture performed in the image sequence. Secondly, by transforming each image first into a sequence of spatially ordered stripes and then producing a VQ label stream by appending each spatial VQ stream of one image at the end of the preceeding frame in the time sequence, a mixture between temporal and spatial ordering of the image stripes has been produced. Such a mixture could not be considered as the best possible feature representation of the video sequence, although it worked already surprisingly well. Since further improvements could have been expected by applying more suitable feature extraction methods, the ultimate goal of such new feature extraction algorithms has been determined by the following requirements:

- to avoid a mixture between spatial and sequential information
- to find features representing the characteristics of the dynamics in the sequence and thus representing the performed gesture
- to be robust against variations of the gesture and of the person performing the gestures
- to be robust against variations of the background
- to drastically reduce the overall data size of the transformed sequence while keeping the information necessary for identification of the gesture

Especially the last point is very important for any effective image sequence recognition algorithm. Since the amount of data to be processed is already very high for most static image processing applications, it becomes extremely high for any image sequence processing algorithm. It is therefore the difficult task of a suitable feature extraction method to reduce that data as much as possible while preserving exactly the information that is urgently needed for performing the recognition task.

As already shown in [1], the velocity movie, derived from the original video sequence by subtracting adjacent frames, has proved to be more suitable for person-independent gesture recognition than the original image sequence. Consequently, this velocity movie was taken as the basis for the new feature extraction method. One frame of this velocity movie contains only the changing parts of the image, i.e. it characterizes the aspects of the motion

performed in this frame, as compared to the previous frame. The size of the grey values d(x,y) in this frame indicates the intensity of the motion in each position (x,y) of the image. If one imagines these grey values as a "mountain area" of elevation d(x,y) at point (x,y), then this mountain area can be considered as a distribution of the movement over the image space in x- and y-direction. Each mountain area will be characteristic for a specific motion, belonging to a certain gesture. If it is possible to characterize this mountain area by certain features, these features should be a good representation for the current motion in the velocity frame. For instance, it may be possible that for a certain gesture with the right hand, the movement will be more concentrated in the upper left corner of the image, resulting in a mountain area with high elevation in this part of the image. One possibility to express this fact is the computation of the center of gravity $\underline{m}^T = [m_x, m_y]$ of the mountain area according to the formulas:

$$m_x = \frac{\sum_x \sum_y d(x,y) \cdot x}{\sum_x \sum_y d(x,y)} \qquad m_y = \frac{\sum_x \sum_y d(x,y) \cdot y}{\sum_x \sum_y d(x,y)} \tag{1}$$

The vector $\underline{m}$ can then be also interpreted as "center of motion" of the image. This center will be certainly shifted towards the upper left part of the image in our example for a right hand gesture and will be characteristic for this specific motion. Another useful feature will be the average deviation of the motion in all points of the image from the center of the motion, defined in horizontal direction as:

$$\sigma_x = \frac{\sum_x \sum_y |d(x,y) \cdot (x - m_x)|}{\sum_x \sum_y |d(x,y)|} \tag{2}$$

and by a similar formula for σ_y in vertical direction. This feature can be very helpful for distinguishing a gesture where large parts of the body is in motion (e.g. a step to the left) from a gesture concentrated more in a smaller area, where only a small body part moves (e.g. hand waving). This feature can be also considered as "wideness of the movement". As shown in [3], the use of $\underline{m}$, σ_x and σ_y also results in an interesting way to visually represent a motion in the image as an ellipsis centered in $\underline{m}$ and with σ_x and σ_y as main axes. For instance, an image with a flat small ellipsis in the upper left hand corner might characterize a horizontal movement of the elevated right hand, e.g. as one instant of a hand waving gesture. Another important feature that can be included in this representation is the "overall intensity of motion" which is simply the average height of the absolute mountain area, expressed as

$$i = \frac{\sum_x \sum_y |d(x,y)|}{\sum_x \sum_y 1} \tag{3}$$

where a large value of i represents a very intensive motion of large parts of the body, and a small value characterizes an almost stationary image. If this value is added as grey value to the previously mentioned ellipsis, then an image sequence can be visually represented as shown in Fig. 1.

The upper part of Fig. 1 shows a sequence of velocity frames with the resulting ellipsis overlayed in each frame. If this overlay is performed in the original sequence, a very insightful feature representation of the actual gesture can be obtained, as shown in the lower part of Fig. 1. It can be seen that the gesture "hand waving both" results in a mostly flat and centered ellipsis, because the motion is always equally distributed between the upper left and right parts of each frame. The wider the hands are apart, the longer the ellipsis becomes due to a large value of σ_x, indicating that the actual motion is far away from its center of gravity. The changing grey value of the ellipsis (better visible in the upper part of Fig. 1) indicates if the frame is shot during full motion (resulting in a darker ellipsis) or at a point where the hands have just reached a maximum horizontal amplitude and begin to change direction.

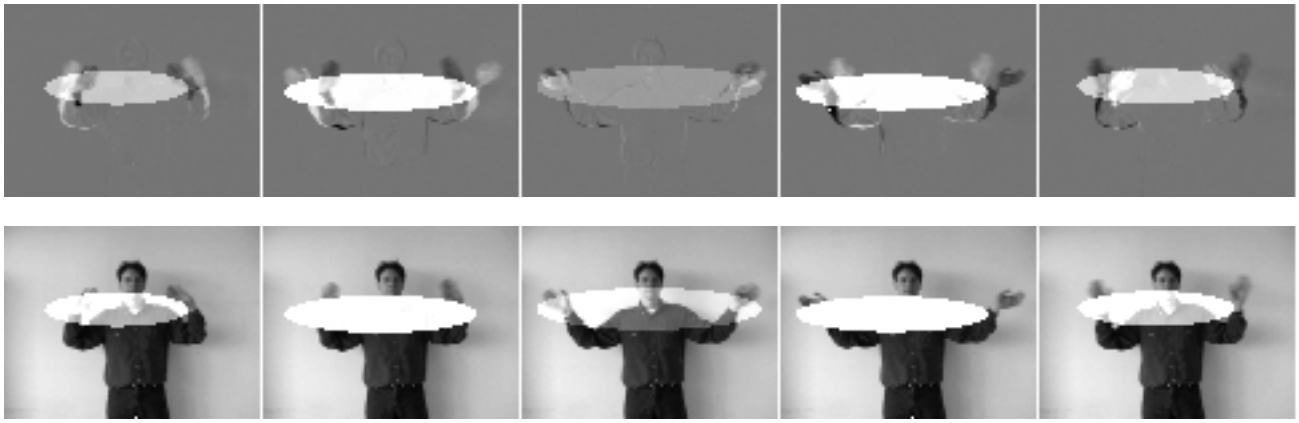

Fig.1: Velocity movie (upper part) and original video sequence (lower part) with overlayed motion ellipsis

Additionally to the 5 major features m_x, m_y, σ_x, σ_y, and i, some more features can be computed, e.g. by separating the parts of the mountain area containing negative values for d(x,y) - which can result from smaller grey values of the previous frame in this area - from all positive areas. It can be shown, that such considerations are helpful for distinguishing certain confusable gestures, e.g. "nod-yes" and "nod-no". Basically, it is possible to compute all 5 features mentioned before also separately only for all positive values of d(x,y) and all negative values. In practice, it has turned out that only the separate calculation of the motion centers according to Eq. (1) is helpful for distinction of difficult to separate gestures. This leads finally to a feature vector consisting of 7 different features, 5 of them as just mentioned before and shown in the ellipsis in Fig. 1. The remaining 2 features consist of the distance between motion centers of positive and negative values in horizontal and vertical direction, respectively. It should be noted, that this feature extraction method represents a remarkable reduction in complexity and dimension, by scaling down an image of 96x72=6912 grey values into a 7-dimensional vector, while preserving the characteristics of the currently observed motion. This 7-dimensional motion vector is derived for each frame, resulting in a vector sequence, where each vector carries important information about the current motion, and thus the entire sequence contains the information about the performed gesture.

3. RECOGNITION METHOD AND RESULTS

Basically, it would have been possible to use for the new features exactly the same HMM structure as described in [1] for the gesture recognition system originally developed for testing the statistical image sequence recognition algorithms. In the original discrete system, the spatial vectors presented to the vector quantizer where of large dimension and several of such vectors have been used to represent one single frame of the video sequence. Therefore, it would not have been very feasible to model such vectors directly by a continuous HMM. However, it should be noted that the new feature vector represents an entire single frame and that it is a vector of continuous values which can be easily modeled by continuous density HMM's, e.g. using Gaussian mixtures for pdf modeling. It was therefore decided to test the suitability of continuous mixture density HMM's for this task in combination with the new reduced feature vectors, and thus such continuous HMM's were used to train and test the system on an extended version of the image database used already in [1]. It turned out that continuous HMM's achieved a higher recognition rate than discrete HMM's for a task with the same training and test data using the new feature extraction method in both cases. Only the new feature extraction method enabled the use of continuous HMM's, both in terms of keeping the number of parameters in a certain limit and in order to be able to maintain the real-time capability of the system while switching from discrete to continuous HMM's.

While the original system with old spatial feature extraction and discrete HMM's achieved a recognition rate of 90,0% on a 10-gesture recognition task similar to Task 1 in Table 1, the new system obtained 96,7% on this same task, corresponding to a remarkable 67% relative error reduction. On the new defined Task 1, as shown in Table 1, the new system obtains even a recognition rate of 99,5%, because the 12 gestures there can be slightly better separated than the 10 gestures used in the old system. Even for the very complex 24-gesture recognition task defined in Task 2, the new system obtains 91,7% recognition rate, which we consider as very high for such a demanding task.

4. CONCLUSION

A new feature extracion method for image sequence recognition based on a statistical approach using hidden Markov models has been presented. The use of the new features has resulted in the development of a new version of our high performance gesture recognition system with an improved architecture and real-time capabilities. We believe that this approach for image sequence recognition has already resulted in a very powerful and robust prototype system for gesture recognition. We also believe that this approach will open up further possibilities for improvements, in feature extraction as well as in stochastic visual modeling, or in discriminative training techniques for video sequences, and we will continue to further improve our system in these directions. Future research is concentrating in 2 directions: The first is the improvement of the robustness of the system against displacements and rotations of the person performing the gestures, and the second is the use of the statistical image sequence recognition algorithms for content-based video indexing.

5. REFERENCES

[1] M. Schuster, G. Rigoll, "Fast Online Video Image Sequence Recognition with Statistical Methods," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, 1996, pp. 3450-3453

[2] G. Rigoll, A. Kosmala, M. Schuster, "A New Approach to Video Sequence Recognition Based on Statistical Methods," Proc. IEEE Int. Conf. on Image Processing (ICIP), Lausanne, 1996, Vol. III, pp. 839-842

[3] S. Eickeler, "Optimization of a System for Recognition of Video Sequences," Diploma Thesis, Gerhard-Mercator-University Duisburg, Faculty of Electrical Engineering, 1996 (in German)