Appearance-Based Person Recognition for Surveillance Applications

Lutz Goldmann, Mustafa Karaman, Jose Tomas Saez Minquez, Thomas Sikora

Technical University of Berlin, Communication Systems Group
Einsteinufer 17, 10587 Berlin, Germany

Abstract This paper presents an original system for recognizing persons based on their appearance. Thus, it is especially suitable to surveillance scenarios, where biometric information might not be available. Different visual low level features in combination with different supervised learning methods are examined in order to built a robust system. Furthermore, complementary features are fused using postmapping fusion concepts to improve the reliability. The experiments show that the system is able to distinguish a large number of people and can be used for different applications.

1 Introduction

Visual surveillance has gained considerable interest recently due to its important role in security. Most research so far has concentrated on detecting and tracking humans and interpreting their behaviour. Nowadays also the recognition of persons in these surveillance scenarios gains more and more interest. By combining tracking and recognition it becomes possible to "relate location to identity".

In general, visual person recognition can be either based on biometric features (face, gait) or non-biometric features (appearance). Biometric features are based on unique characterstics of individual persons and thus offer a high discriminabiltiy. Nevertheless, they also impose strong restrictions on the data which limits their application. Face recognition usually requires high resolution images with frontal faces in order to work reliably while gait recognition is based on full body profile images. On the other hand, non-biometric features lack the uniqueness of biometric ones and usually have a shorter validity period, but they impose much weaker restrictions on the data. Appearance-based features such as color and texture can be extracted under diverse conditions and are widely used for object recognition.

Only very few work [7, 2] has been done in the field of appearance-based person recognition. The main limitations of the systems so far are their restrictions concerning the environment. This is caused by the lack of sophisticated segmentation methods that can cope with typical problems such as shadows and illumination changes. Nakajima [7] restricts the system to an indoor environment and uses a very simple detection approach. Haehnel [2] deals with a much more unrealistic scenario, where a blue screen is used to ease the detection step. This allows to distinguish a higher number of persons since the classification is uneffected from segmentation problems. Furthermore, both restricted their experiments to very complicated features and a small set of classifiers.

We propose an automatic system that can be used in typical environments and is not restricted to laboratory conditions. Different color and texture features are examined and the classification methods are extended to provide a more thorough analysis. Since we use simpler features, our system supports real-time applications. Furthermore, the fusion of complementary features such as color and texture is considered to improve the performance.

2 System overview

Figure 1 gives an overview of the overall system. It consists mainly of video acquisition, person detection, feature extraction, classification, and optionally fusion.

Since a supervised learning approach is used, the system operates in two different modes. During the training, a model for each person is created using ground truth data. Based on these models the classification of unknown persons is carried out in the testing mode.

The system supports two different applications. It can serve as a standard person recognition system in a surveillance environment, where one place (e.g. a room), is monitored using a single camera. In that scenario it supports short term identification of people entering, exiting, and reentering the place. On the other hand, it is

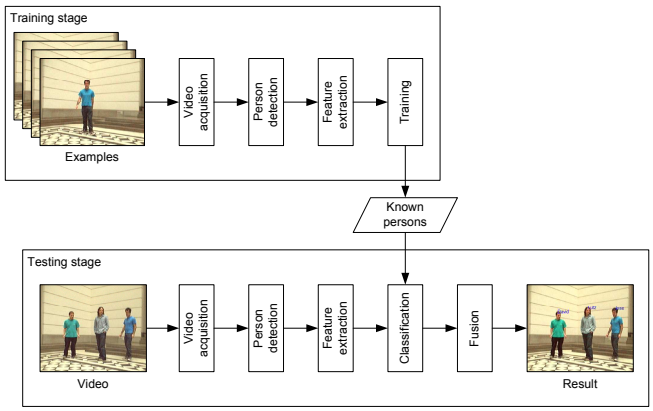

Fig. 1 System overview.

also suitable for camera handover between multiple cameras with non-overlapping views (e.g in corridors). This is especially useful in situations where a person needs to be tracked over a large area with multiple cameras.

2.1 Video acquisition

The first step of each video-based system is the acquisition of the video data. The choice of the camera including sensor type, sensor quality, resolution, and mobility is crucial to the person detection and feature extraction stage. It also influences the performance of the recognition system. Thus, two different types of cameras are considered, a digital camcorder and a webcam. Both cameras are static and utilize color sensors. They mainly differ in the resolution, frame rate, and sensor quality.

2.2 Person detection

Assuming fixed cameras, static background segmentation methods can be used in order to separate interesting foreground objects from the background. The method proposed by Horprasert et al. [3] is adopted, based on a comparison of various state of the art methods [4]. The used method gives the best trade-off between segmentation quality and computational complexity. Horprasert et al. utilize a pixel-based background model including luminance and chrominance information which is trained in advance from multiple background images. For an actual pixel both luminance and chrominance distortion to the background model are calculated. Based on these values each pixel is classified into background, shadow, highlight, or foreground. This multi-class approach allows to cope with several problems such as shadows and highlights.

The resulting binary foreground mask passes through a connected component labeling stage resulting in objects consisting of single connected blobs. Persons are detected by applying heuristic rules to these blobs based on size and shape criteria.

2.3 Feature extraction

Based on the binary object mask and the original image, visual low level features are extracted for describing each of the detected persons. Since the appearance of a person is dominated by its clothes, color and texture features are suitable for the description. The following descriptors are used:

Average RGB value (ARGB): The average RGB value describes the color of an object by averaging each channel (R, G, B) over all pixels that compose the object. The average RGB is inherently invariant against scale, rotation, and translation.

Color structure descriptor (CSD): The color structure descriptor is defined as part of the MPEG-7 visual standard [5]. Unlike color histograms it takes the distribution of the colors in an object into account. This is done by moving a sliding window over the object and determining the colors within the window without considering the actual number of pixels of a certain color. The MPEG-7 standard recommends the HMMD color space and suggests a combined quantization for all color channels into 32 bins which is motivated by research on the human visual system (HVS). Like most of the MPEG-7 visual descriptors it is invariant against scale, rotation, and translation.

HMMD color histogram (CH): Color histograms are widely used for describing the color of objects. They mainly differ in the used color space and the quantization levels. Here, the RGB pixels are converted to HMMD color space and quantized into 32 bins for all channels together [5]. The histogram is extracted by counting the number of pixels for each bin. The resulting 1D histogram is inherently invariant to rotation and translation. In order to make it scale invariant, the histogram is normalized by the number of pixels.

Intensity histogram-based features (IH): The intensity histogram shows (for each intensity level) the number of pixels which have a certain intensity value. Since pixels are considered independently, it contains the first order statistical information about an object. Different features can be calculated by this histogram to characterise textures [6]. Since some of the features can reach undefined values, our system utilizes only the mean, variance, and energy.

Cooccurrence matrix-based features (CM): Another way of describing textures is based on second order statistics and uses cooccurrence matrices. These are square matrices of dimension equal to the number of intensity levels that contain the joint probability values $p_{d,\theta}(i,j)$ of pairs of pixels with certain intensity values i and j for different angles θ and distances d. A reduced number of

features can be calculated based on these cooccurrence matrices [6]. Given that, some of the features can take undefined values and others give very little information only the energy, absolute value, and contrast are used in our system.

2.4 Recognition

Different parametric, non-parametric, and discriminating methods for supervised learning are considered. For a given unknown person, the output of the classifiers are either opinions (probabilities, scores) corresponding to all possible persons in the database or a decision (label) in favor to one person out of the database. The following methods are considered:

k-nearest neighbor (kNN): The kNN belongs to the non-parametric pattern recognition methods. It is a very intuitive method that classifies unlabeled test samples based on their similarity to labelled training samples. For a given unlabeled sample, it finds the k closest samples in the training data set and assigns the class label c_i 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 d is assigned.

Gaussian mixture model (GMM): The GMM is a parametric pattern recognition method. It models the probability density function (PDF) of one class using a linear combination of Gaussian distributions. The parameters of the Gaussian distributions and the prior probabilites used for the linear combination are calculated during the training stage based on the expectation maximization (EM) algorithm. For multi-class problems one GMM is trained for each class c_i. Given an unknown sample x the likelihood probabilities $p(x|c_i)$ for each class are calculated and the decision is made using the maximum a posteriori (MAP) criterion.

Support vector machine (SVM): The SVM is another non-parametric classification method that separates two classes using an optimal separating hyperplane. This hyperplane is found by maximizing the margin between two classes using support vectors. Since a linear separation is often impossible, it operates in two stages. First, a kernel function is used to map the feature vector non-linearly into a high-dimensional space and then an optimal separating hyperplane is constructed within this space. Although SVMs are inherently classifiers for two class problems, they can be extended to multi-class problems using different structures of multiple SVMs.

2.5 Fusion

Information fusion [8] in general deals with the combination of different sources of information in order to utilize their complementarity to improve the systems performance. In the system post mapping fusion is considered by combining the output of different classifiers with each other. Based on the output data, opinion or decision level fusion is applied. Opinion level fusion techniques include weighted summation, weighted product, median, and max rule. For decision level fusion majority voting, ranked lists, AND, and OR fusion can be used.

Descriptor	GMM	kNN	SVM
Average RGB	90.1	86.5	90.2
Color histogram	90.3	94.7	89.7
Color structure descriptor	90.6	94.5	90.3
Intensity histogram	59.1	55.2	39.4
Co-occurence matrix	24.8	23.7	27.1

Table 1 Recognition rates (%) for single descriptors.

3 Experiments

Several experiments were conducted in order to obtain reliable results concerning the optimal choice of descriptors, classifiers, and fusion techniques. The evaluation of the person recognition system was based on manual annotated ground truth data and confusion matrix derived measures such as recognition rate (RR), true positive rate (TPR), false positive rate (FPR). Furthermore, receiver operating characteristic (ROC) curves were used to analyze the performance of detecting unknown persons.

An own database was built due to the lack of suitable public available databases. It consists of videos containing multiple person (2-10) in different environments (indoor, outdoor). The videos are in CIF resolution and 24bit RGB color space.

Single descriptors: The first experiments focus on single descriptors in combination with different classifiers. For each classifier different parameters sets were evaluated, but only the best results are reported in table 1.

Concerning the classifiers no general result exists. The ranking depends largely on the combination with a certain descriptor. Although the GMM is not the best classifier for all descriptors, it is the most stable classifier over all experiments. Therefore and because it inherently supports the opinion fusion by providing propabilities for each category, it is used throughout the subsequent experiments.

With respect to the descriptors, we observe that color descriptors are generally better than texture descriptors. Reasons for that are the low video resolution and the lack of different textures in the clothes of the persons. In general, the performance of the color descriptors is very similar. The best color descriptor is the color structure descriptor, followed by the color histogram and the average RGB. But if the computational complexity is also

Fusion	Method	GMM
Single descriptor	Average RGB	90.1
	Intensity histogram	59.1
Multi descriptor	Product method	91.8
	Sum method	74.2
	Median method	74.2
	Max method	73.9

Table 2 Recognition rates (%) for multiple descriptors.

considered, the average RGB gives obviously the best tradeoff. The performance of the texture descriptors is rather low and there is huge difference between the intensity histogram and the cooccurence matrix. Thus, if texture descriptors are considered, the intensity histogram will be used.

Multiple descriptors: Based on the ideas of information fusion and after analyzing the confusion matrices of the different single descriptor experiments, it was concluded that the fusion of color and texture descriptors can improve the results due to their complementarity.

Based on the preliminary experiments, the opinions of average RGB and intensity histogram in combination with GMMs are fused. Different opinion level methods as mentioned in section 2.5 are considered. The results are shown in table 2.

It illustrates that better performance can be achieved by the combination of the two descriptors. The recognition rate increases by around 2% from the single descriptor (SD) system (90.1% and 59,1%) to the multi descriptor (MD) system (91.8%).

Unknown person detection: All the preceding experiments are based on the closed set scenario [1], which means that only known persons are to be identified. If the person may or may not be known, it is called open set scenario and the system must decide if the person is known or unknown before identifying it. This is usually based on thresholding the maximum probability of the recognition stage. The threshold can be adjusted depending on the application.

Different experiments were conducted, where a single person was excluded from the training stage and treated as unknown person during the testing stage. Figure 2 shows typical ROC curves for a selection of different persons and the worst case.

The curves vary for different persons, which is caused by the visual similarity of the unknown person with one or more known persons. If the visual appearance of the unknown person is highly similar with that of a known person, the unknown person detection will produce higher false positive and false negative rates which will result in a less optimal ROC curve. This can be seen for Lars whose ROC curve is the worst compared to the others. In contrast Jorge leads to a nearly optimal ROC curve.

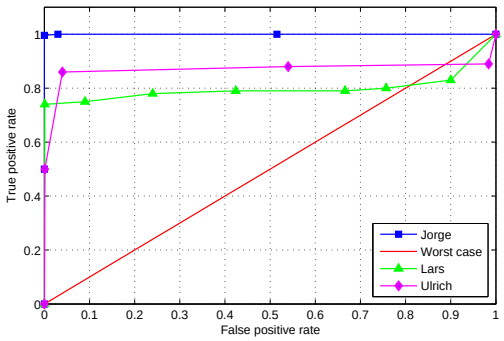

Fig. 2 ROC curve for unknown person detection.

4 Conclusion

The results show that appearance-based person recognition can provide reliable results for a surveillance scenario. Obviously, the number of discriminable persons is smaller and the validity period shorter than in face recognition. Nevertheless, for a limited time and a smaller number of people it can be used for both identification and camera handover.

Acknowledgement: The research leading to this paper was supported by the European Commission under contract FP6-027026, Knowledge space of semantic inference for automatic annotation and retrieval of multimedia content – K-Space.

References

1. P. Angkititrakul and J. H. L. Hansen. Identifying in-set and out-of-set speakers using neighborhood information. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.
2. M. Haehnel, D. Kluender, and K.-F. Kraiss. Color and texture features for person recognition. *International Joint Conference on Neural Networks (IJCNN)*, Jul 2004.
3. T. Horprasert and D. Harwood. A statistical approach for real-time robust background subtraction and shadow detection. Technical report, University of Maryland, 1999.
4. M. Karaman, L. Goldmann, D. Yu, and T. Sikora. Comparison of static background segmentation methods. In *Visual Communications and Image Processing (VCIP)*, 2005.
5. B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7*. John Wiley & Sons Ltd., 2002.
6. A. Materka and M. Strzelecki. Texture analysis methods – a review. COST B11 report, University of Lodz, 1998.
7. C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full body person recognition system. *Pattern Recognition*, 2003.
8. C. Sanderson. *Automatic Person Verification Using Speech and Face Information*. PhD thesis, Griffith University, 2002.