PERSON IDENTIFICATION IN SURVEILLANCE VIDEO BY COMBINING MPEG-7 EXPERTS[1]

B. Birant Örten[2], Medeni Soysal[2,3] and A. Aydın Alatan[2,3]
[2]Department of Electrical and Electronics Engineering, M.E.T.U.
[3]TÜBİTAK BİLTEN,
Balgat 06531 Ankara TURKEY

ABSTRACT

Identification of people in surveillance videos is an important problem and MPEG-7 visual descriptors are utilized for such recognition in a regional manner, which result from independently moving subjects in front of stationary cameras. While background modeling is achieved by using a hierarchical non-parametric Parzen-window approach, the resulting regional descriptors are classified by combining experts via different combination rules. Simulation results enjoy a promising recognition performance for the tested data set.

1. INTRODUCTION

Automation is the sole answer for the increased demand for personal and societal security in daily life. Since off-the-shelf cameras become vastly available, the automatic analysis of such huge amount remains as a major challenge.

The main focus of this paper is to examine the performance of an object-based video retrieval system, which is utilized for surveillance applications. In this system, MPEG-7 visual descriptors are tested only within segmented regions, which are assumed to result from independently moving subjects in front of stationary cameras.

2. MOVING OBJECT DETECTION IN SURVEILLANCE SYSTEMS

The performance of an automated visual surveillance system considerably depends on its ability to detect moving objects in the observed environment. A subsequent action like tracking, analyzing the motion or identifying persons requires an accurate extraction of the foreground objects, making moving object detection a crucial part of the system.

2.1. Related Work

Extensive research effort has been devoted to moving object segmentation from video imagery. In the literature, there are basically two conventional methods: *temporal differencing* and *background modeling and subtraction*. The former approach is possibly the simplest one, also capable of adapting to changes in the scene with a low computational load. However, the detection performance of temporal differencing is usually quite poor in real-life surveillance applications. On the other hand, background modeling and subtraction approach has been used successfully in several algorithms in the literature.

Haritaoglu *et al.* [4], models the background by representing each pixel with its maximum intensity value, minimum intensity value and intensity difference values between consecutive pixels. The limitation of the model is that it is not very robust under illumination changes.

Oliver *et al.* [7] have proposed an eigenspace model for moving object segmentation. In this method, dimensionality of the space constructed from sample images is reduced using Principal Component Analysis (PCA). The claim is that, after the PCA, reduced space will represent only the static parts of the scene, yielding moving objects if an image is projected on this space. Although the method has some success in certain applications, it cannot model dynamic scenes well. Hence, it is not very suitable for outdoor surveillance tasks.

Another statistical method is proposed by Wren, *et al.* [3], which models every point in the scene using a Gaussian with a mean color value and a distribution around it. The drawback of the model is that it can only handle unimodal distributions. Later, in a general approach, mixture of Gaussians is also used instead of a single Gaussian [9].

Elgammal, *et al.* [1] use sample background images to estimate the probability of observing pixel intensity values in a nonparametric way. As a matter of fact, this method is theoretically well established and yields much accurate results under challenging outdoor conditions.

2.2. Hierarchical Parzen Window-based Moving Object Detection

In this section, the utilized method to model the background, which is a hierarchical version of [1], is described. This approach depends on nonparametrically estimating the probability of observing pixel intensity values based on the sample intensities. An estimate of the pixel intensity can be obtained using,

$$p(x) = \frac{1}{N}\sum_{k}\varphi(x - x_k) \qquad (1)$$

where the set $\{x_1, x_2, \ldots, x_N\}$ gives the sample intensity values in the temporal history of a particular pixel in the image. Function $\varphi(.)$ in (1) is the window function, which is used for interpolation, giving a measure for the contribution of each sample in the estimate of $p(x)$. When the window function is chosen as Gaussian, (1) becomes:

[1] This work is supported by DPT within project 2004K120720 and TÜBİTAK under COST 292.

$$p(x) = \frac{1}{N}\sum_{k}\prod_{i=1}^{3}\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{\frac{-(x_i - x_{ki})^2}{2\sigma_i}} \qquad (2)$$

Above equation can be obtained for three color channels (*R, G, B*) using the assumption that they are all independent, where σ_i is the window function width of the i^{th} color channel window function. Considering the samples $\{x_1, x_2, \ldots, x_N\}$ are background scene intensities, one can decide whether a pixel will be classified as foreground or background according to the value of (2). This process yields the first stage detection of objects.

To improve the results, a second stage should also be used. At this stage, using the sample history of neighbors of a pixel (instead of its own history values), probability that a pixel belongs to the background is calculated. This approach reduces false alarms due to dynamic scene effects, such as tree branches or a flag waving in the wind. Another feature of the second stage is the connected component probability estimation. This process yields, whether a connected component is displaced from the background or it is an appeared object in the scene. The second stage helps reducing false alarms in a dynamic environment providing a robust model for moving object detection.

Although the above-mentioned method is effective for background modeling, it is slow due to calculations at the estimation stage. Hence, one should utilize multi-level processing to make the system appropriate for real-time surveillance applications.

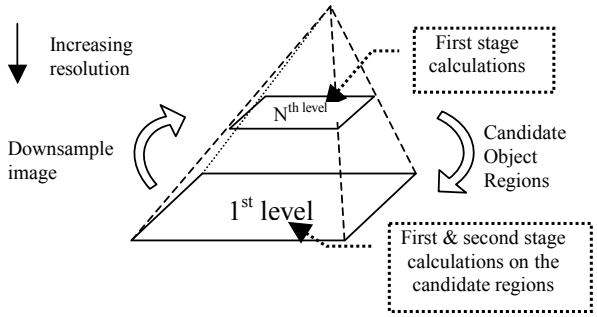

Fig. 1. Hierarchical detection of moving objects

Figure 1 illustrates the hierarchical structure of the proposed system. A frame from the sequence is downsampled and first stage detection is performed on this low-resolution image. Because of the high detection performance of the nonparametric model, the object regions are captured quite accurately even in the downsampled image, providing object *bounding boxes* to the upper level. Hence, upper level calculations are only performed on the candidate regions instead of whole image, ensuring faster detection performance. Indeed, processing the whole frame in test sequence takes approximately 5 sec. whereas hierarchical system makes it possible to process the same frame around 150-200 msecs. Besides, providing a bounding box to the upper level only makes the processing faster without causing any performance change in the final result.

3. PERSON IDENTIFICATION WITH REGION-BASED MPEG-7 DESCRIPTORS

When classifying people in the surveillance videos, it is assumed that color and texture are the most important and invariant visual features. Obviously in different applications, shape descriptors might also be used to discriminate between human, animal and vehicle classes. Color structure and homogeneous texture descriptors of MPEG-7 standard [2] are selected to represent these features based on some past experience [11].

3.1. Utilized MPEG-7 Descriptors

A brief explanation is presented for the MPEG-7 descriptors, which are used for person identification:

Color Structure: MPEG-7 Color Structure descriptor is used in the experiments to represent the color feature of an image. Color Structure descriptor specifies both color content (like color histogram) and the structure of this content by the help of a structure element [2]. This descriptor can distinguish between two images in which a given color is present in identical amounts, whereas the structure of the group of pixels is different. During simulations, 64-bin version of Color Structure descriptor is utilized.

Homogeneous Texture: The second basic feature of an image, texture, is represented by MPEG-7 Homogeneous Texture descriptor, characterizing the region texture by mean energy and energy deviation from a set of frequency channels. The definition of this descriptor in MPEG7 standard permits its use on arbitrary shaped regions.

The channels are modeled by Gabor functions and the 2-D frequency plane is portioned into 30 channels. In order to construct this descriptor, mean and standard deviation of the image in pixel domain is calculated and combined into a feature vector with the means and energy deviations computed in each of the 30 frequency channels. As a result, a feature vector of 62 dimensions is extracted from each image [2].

4. EXPERT COMBINATION FOR CLASSIFICATION

4.1. Definition of Experts

Experts are defined as the instances of classifiers with distinct natures or working on distinct feature spaces. In this research, both experts have Support Vector Machine (SVM) [8] nature. SVM performs classification between two classes by finding a decision surface via certain points of the training set. This approach is different in a way that it handles the risk concept. Although other classical classifiers try to classify training sets with minimal errors, SVM can sacrifice from training set performance for being successful on yet-to-be-seen samples [8]. Briefly, one can say that SVM constructs a decision surface between samples of two classes, maximizing the margin between them. SVM classifies test data by calculating the distance of samples from the decision surface with its sign signifying which side of the surface they reside.

On the other hand, in order to combine the classifier outputs, each classifier should produce calibrated posterior probability values. In order to obtain such an output, a simple logistic link function method, proposed by Wahba [6]

$$\mathrm{P(in\text{-}class} \mid \mathrm{x}) = \frac{1}{(1+\exp(-f(x)))} \qquad (3)$$

is utilized. In this formula, $f(x)$ is the output of a SVM, which is the distance of the input vector from the decision surface.

These two SVM experts works on Color Structure and Homogeneous Texture features that are described in Section 3.1.

4.2. Combining Experts

Combining experts have been a popular research topic for years. Latest studies have provided mature and satisfying methods [5]. In this research, 7 different combination methods are evaluated. The first method is *Sum rule*, which in the two expert case simplifies into an arithmetic average of the two probabilities. *Product rule* is another well-known method [5], which is specified by the following formula,

$$P(p_x \mid x, y) = \frac{P_1(p_k \mid x) \times P_2(p_k \mid y)}{P_1(p_k \mid x) \times P_2(p_k \mid y) + (1 - P_1(p_k \mid x)) \times (1 - P_2(p_k \mid y))}$$

where $P_1(p_k|x)$ and $P_2(p_k|y)$ are the single expert probabilities of a person being p_k according to the features x and y. *Max rule* is represented by a similar formula.

$$P(p_x \mid x, y) = \frac{Max(P_1(p_k \mid x), P_2(p_k \mid y))}{Max(P_1(p_k \mid x), P_2(p_k \mid y)) + Max((1 - P_1(p_k \mid x)), (1 - P_2(p_k \mid y)))}$$

Min rule is in the same format with the max rule except taking minimum values instead. Calculating the geometric mean of the probabilities is another method and gives results that are similar to the product rule. Lastly, *absolute max* and *absolute min* rules are used as special cases of majority vote rule. Absolute max rule picks the highest of the probabilities as the last decision, while absolute min rule picks the lowest probability and assigns this to the sample. Figure 2 illustrates the use of experts in a combination scheme.

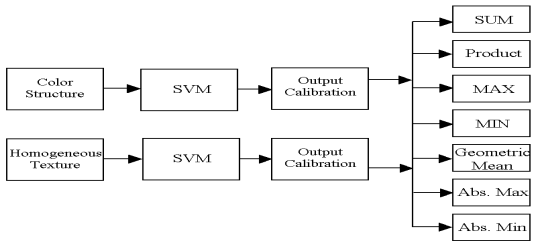

Fig.2. Expert Combination Schemes

5. SIMULATIONS

In this section, the simulation results for the moving object detection and person identification are presented.

5.1. Simulation Set-Up

One of the sequences used in this paper is obtained from MPEG-7 Test Set, (CD# 30, ETRI Surveillance Video), in MPEG-1 format taken at 30 fr/s with a resolution 352x240. The other sequence can be downloaded from the below link (MPEG-1 format, 30 fr/s with a resolution 320x240). http://www.cs.rutgers.edu/~elgammal/ Research/BGS.

5.2. Simulation Results for Moving Object Detection

In this section, the simulation results for moving object detection is presented and discussed. For each video, a comparison of the following algorithm outputs is shown: moving average, eigen-background [7] and proposed hierarchical Parzen windowing.

In Figure 3, a sample frame from ETRI Surveillance video is given together with the outputs of three algorithms. The results for eigenbackground and hierarchical parzen window methods are both satisfactory, whereas moving average produces a ghost-like replica behind the object due to its use of very recent image samples to construct a reference background frame.

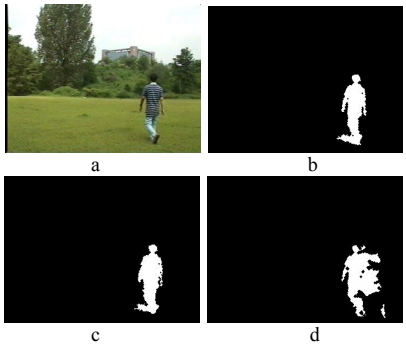

Fig. 3. (a) Original image (b) Eigenbackground (c) Hierarchical parzen windowing (d) Moving average

Another video (Fig.4) contains a dynamic background due to dense tree leaves and branches waving in the wind. The proposed model extracts the object silhouette successfully. However, moving average and eigenbackground approaches yield noisy and inaccurate outputs. Obviously, noise filtering or morphological operations can be used to improve the results of these two methods at the risk of distorting object shape.

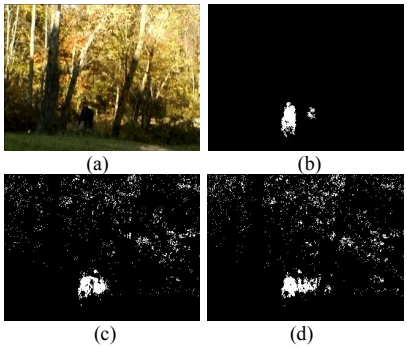

Fig. 4. (a) Original image (b) Hierarchical parzen windowing (c) Eigenbackground (d) Moving average

5.3. Simulation Results for Person Identification

In classification simulations, two persons from ETRI video are identified out of seven distinct identities. Typical examples of frames, containing these distinct identities, are given in Fig. 5. Equal-sized train and test sets are used for constructing one-against-all classification scheme for identifying the two people that are above others in the figure.

In addition to the decisions of single experts based on color and texture, seven different expert combination results are also

		Single Expert		Combined Experts						
		Color	Texture	Sum	Product	Max	Min	Geo. Mean	**Abs Max**	Abs Min
Person 1	Accuracy	92.94%	44.77%	67.64%	67.64%	67.64%	67.64%	62.04%	**85.64%**	52.07%
	Precision	100.00%	0.00%	95.71%	95.71%	95.71%	95.71%	97.67%	**84.85%**	100.00%
	Recall	85.28%	0.00%	34.01%	34.01%	34.01%	34.01%	21.32%	**85.28%**	0.00%
Person 2	Accuracy	66.15%	88.82%	90.37%	90.37%	90.37%	90.37%	85.40%	**90.06%**	64.91%
	Precision	100.00%	89.84%	100.00%	100.00%	100.00%	100.00%	100.00%	**90.00%**	100.00%
	Recall	53.02%	95.26%	86.64%	86.64%	86.64%	86.64%	79.74%	**96.98%**	51.29%

Table 1. Person identification performance results

evaluated. The performance results are given in Table 1.

As can be seen from Table 1, color and texture features of the person to be recognized affects the performance of single expert significantly. For example, in person-1 of Fig.5, color-based expert outperforms the texture-based one, since the texture of the clothes of the person to be identified have no significant difference from the other identities. On the other hand, for the person-2 case, texture-based expert yields better results. Such a problem seems to be solved by combining these experts in an appropriate scheme.

Fig. 5. (Above) Two persons to be identified, person-1 (left) and person-2 (right). (Below) Five Other identities.

In the experiments, it is observed that neither of the color or texture based expert performs well in all cases. Hence, it is by intuition to combine them to reach more stable and successful results. However, in many combinations, the inferior expert degrades the overall performance significantly and hence, the other expert cannot compensate for it. According to Table 1, it can be observed that absolute max combination gives the most stable and promising results. This is due to the nature of this combination, which fits well to situations like this one, where recall is of utter importance and precision is already high enough. The combination strategy of absolute max method prevents the decision to be robust against factors decreasing the recall performance.

6. CONCLUSIONS

Moving object detection is a crucial step in surveillance applications. Parzen window approach proved to be accurate and satisfactory, considering the simulation results. A novel a multi-level analysis stage is also introduced and a considerable speed up is obtained for the test sequences. Attained speed gain is 10-25 times over the original method depending on the object size.

As for the identification part, the seperability of color and texture features of samples varies greatly even in a single domain. Combining experts trained with different features helps to handle this problem as already observed in the previous work [10,11]. For the cases, where recall is relatively important, rigid combination rules, such as product or geometric average become useless, since one expert can easily inhibit the overall decision by giving a low probability result. For such cases, typical in person identification, some other rules, such as absolute max rule should be preferred.

7. REFERENCES

[1] A. Elgammal, D. Harwood, and L. S. Davis. "Non-parametric Model for Background Subtraction." In *Proc. IEEE ICCV'99 FRAME-RATE Workshop*, 1999.

[2] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7 Multimedia Content Description Interface*. John Wiley & Sons Ltd., England, 2002.

[3] C.R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780-785, July 1997.

[4] Haritaoglu, I., D. Harwood and L.S. Davis, "W4: A Real-Time System for Detecting and Tracking People in 2 ½ D." in *5th European Conference on Computer Vision.* 1998. Freiburg, Germany: Springer.

[5] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, March 1998.

[6] J.C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, 1999.

[7] Oliver, N., B. Rosario, and A. Pentland. "A Bayesian Computer Vision System for Modeling Human Interactions." in *Int'l Conf. on Vision Systems*. 1999. Gran Canaria, Spain: Springer.

[8] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[9] W. E. L. Grimson and C. Stauffer, "Adaptive background mixture models for real-time tracking." in *Proc. IEEE Conf. CVPR*, vol. 1, 1999, pp 22-29.

[10] M. Soysal and A. A. Alatan. Combining MPEG-7 Based Visual Experts For Reaching Semantics. International Workshop VLBV03, Madrid, Spain, 18-19 September 2003.

[11] M. Soysal and A. A. Alatan. Combining Region-based MPEG-7 Experts for Reaching Semantics. 6th COST 276 Workshop, Thessaloniki, Greece, May 6-7, 2004.