Proceedings of the 11th International IEEE
Conference on Intelligent Transportation Systems
Beijing, China, October 12-15, 2008

Detection, Tracking and Recognition of Traffic Signs from Video Input

Andrzej Ruta Yongmin Li Xiaohui Liu
School of Information Systems, Computing and Mathematics
Brunel University
Uxbridge, Middlesex UB8 3PH, UK
{Andrzej.Ruta, Yongmin.Li, Xiaohui.Liu}@brunel.ac.uk

Abstract— **In this paper a comprehensive approach to the recognition of traffic signs from video input is proposed. A trained attentive classifier cascade is used to scan the scene in order to quickly establish regions of interest (ROI). Sign candidates within ROIs are captured by detecting the instances of equiangular polygons using a Hough Transform-style shape detector. To ensure a stable tracking of the likely traffic signs, especially in cluttered background, we propose a Pixel Relevance Model, where the pixel relevance is defined as a confidence measure for a pixel being part of a sign's contour. The relevance of the hypothesized contour pixels is updated dynamically within a small search region maintained by a Kalman Filter, which ensures faster computation. Gradient magnitude is used as an observable evidence for this update process. In the classification stage, a temporally integrated template matching technique based on the class-specific discriminative local region representation of an image is adopted. We have evaluated the proposed approach on a large database of 135 traffic signs and numerous real traffic video sequences. A recognition accuracy of over 93% in near real-time has been achieved.**

I. INTRODUCTION

Detection and recognition of traffic signs from video input has long been in the center of interest for having great impact on the safety of the driver. However, it remains a challenging problem. First, it is difficult to track an object in an image sequence captured from a vehicle undergoing a generally non-uniform motion. Second, we have to disambiguate between the true signs and a potentially large number of other natural or man-made objects of similar shape and appearance. Third, there are many road signs and some of them are very similar to one another. Finally, in a realistic scenario a traffic sign recognition system must operate in real time.

Most commonly, a two-stage approach to the detection and recognition of road signs has been adopted in the state-of-the-art literature. A prior knowledge about the signs is incorporated in the detection part in some sort of constraints used to 1) define how to pre-segment the scene in order to find the interest regions, and 2) define the acceptable appearance of signs and the geometrical relationships between their parts with respect to color and shape [3], [4], [6], [9], [10]. The major drawback of this approach is the sequential nature of the processing which makes it impossible to recover from the errors made early in the processing pipeline. Furthermore, many of these studies present heuristics and therefore lack the solid theoretical foundations. Bahlmann et al. [1] have taken a different strategy in which color-parametrized Haar-wavelet shape descriptors are combined within a trainable attentive cascade of classifiers in order to locate road sign candidates in the scene. This method seems to work reasonably accurately for a relatively limited number of signs to recognize and easy traffic scenes. Several studies address the problem of sign tracking over time [4], [7], [10]. However, these attempts are mainly focused on the geometrical tracking, used to reduce the uncertainty on the candidate's position and scale in the image, and are based on the strong motion assumptions, e.g. constant velocity. In the classification stage a pixel-based approach is often adopted and the class of a detected sign is determined by the cross-correlation template matching [10] or neural networks [3]. Feature-based approach is used for instance in [9], where the classification problem is decomposed in a decision tree reflecting the natural grouping of the known road signs. For each subgroup a Laplace kernel classifier is used to classify an unknown sign represented by its various numerical characteristics e.g. moments. In [8], [12] a different strategy was developed based on the idea of representing a candidate sign as a set of similarities to the stored prototype images, each assessed with respect to a class-specific set of local regions refined in the training process.

In this paper we present an innovative algorithm for fast detection, tracking and classification of traffic signs from a moving vehicle. An attentive classifier cascade is adopted to minimize the search region at a very early stage and hence reduce computation. Further, we introduce a confidence measure of an individual pixel being part of a sign's contour. Based on this notion we build a contour tracking framework using a spatio-temporal voting scheme. The proposed method is efficiently integrated with a Kalman Filter which is used, apart from to reduce the search region, also as a regularizer of the relevance model. As a result, the true sign's contours become easier to detect as the undesirable edges that are not a part of this contour are suppressed. To facilitate this, we extend the equiangular polygon detection algorithm of Loy and Barnes [6] by making it operate on the abovementioned pixel relevance. Therefore, the detector attempts to capture the contour of a sign being tracked in a feature image that incorporates both the current-frame gradient information and the past observations. In the classification stage we propose to represent each road sign by a compact, trainable subset of its most discriminative local regions i.e. these in which this sign looks possibly the most different from all other signs. Temporally integrated template matching based on this class-

1-4244-2112-1/08/$20.00 ©2008 IEEE

specific representation is introduced and evaluated through the experiments on the traffic video.

The rest of this paper is organized as follows: In section II the concept of pixel relevance and its realization in detecting signs are discussed. Section III describes a discriminative local region representation of road signs and a dynamic classifier used for recognition of the observed candidates. Section IV presents experimental results of detection and recognition on the real traffic video sequences. Finally, conclusions are drawn in section V.

II. SIGN DETECTION AND TRACKING

A. *Pixel Relevance Model*

We start with a definition of pixel relevance. We call a pixel x_{ij} relevant if it belongs to the contour of a road sign and irrelevant otherwise. Formally, relevance r_{ij} of pixel x_{ij} is defined as a real number between 0 and 1, i.e. x_{ij} is completely irrelevant when $r_{ij} = 0$ and completely relevant when $r_{ij} = 1$. As the signs move through the scene and grow in size while being approached by the vehicle, relevance must change over time as well. We model the dynamics of this process using a spatio-temporal voting graph, fragment of which is shown in Fig. 1. The graph encapsulates the relevance distribution over an entire image region at time t, $\mathbf{r}(t)$, in a set of state nodes. Evolution of a single pixel's relevance is assumed to be a first order stationary Markov process and the supposingly weak correlations between the same-slice pixels are ignored. Relevance of each pixel x_{ij} at time t, $r_{ij}(t)$, is dependent on the relevance of its neighborhood in the previous frame, $r_{N(i,j)}(t-1) = \frac{1}{n}\sum_{x_{kl}\in N(i,j)} r_{kl}(t-1)$ (where n is the size of the neighborhood), and the observable feature at time t, $f(x_{ij}(t))$.

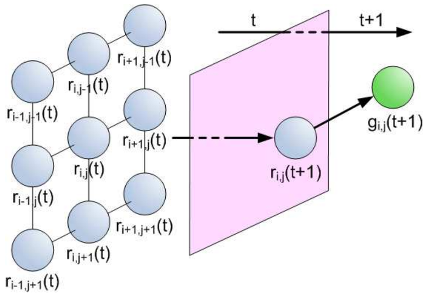

Fig. 1. Fragment of the spatio-temporal voting structure used to model the dynamics of pixel relevance. Consecutive time slices are shown.

Our state transition model is defined by the following function supported on a $[0, 1]$ interval:

$$\phi_{ij}(t) \sim \left(1 - e^{-k_T r_{N(i,j)}(t-1)}\right) \qquad (1)$$

for some constant k_T. We are postponing the more precise definition of the transition function to section II C. Relevance projected from time slice t to $t+1$ is further conditioned on the observed feature at time $t+1$, and we define this update process by the same class of functions:

$$\psi_{ij}(t) = 1 - e^{-k_O f(x_{ij}(t))} \qquad (2)$$

for another parameter k_O. Using the above definitions, evolution of the pixel relevance can be expressed as:

$$\begin{array}{l} r^-_{ij}(t+1) = r_{N(i,j)}(t)\phi_{ij}(t+1) \\ r^+_{ij}(t+1) = r^-_{ij}(t+1)\psi_{ij}(t+1) \end{array} . \qquad (3)$$

In sections II B-C we will discuss how the Pixel Relevance Model is integrated with sign detection and tracking.

B. *Detection*

Our road sign detector is triggered every fixed number of frames to capture new candidates emerging in the scene. It makes use of the *a priori* knowledge about the model signs, uniquely identified by their general shape, color and contained ideogram. Based on the first two properties four European sign categories coinciding with the well-known semantic families are identified: instruction (blue circular), prohibitive (red circular), cautionary (yellow triangular), and informative (blue square) signs. The proposed detector operates on the edge and gradient maps of the original video frames extracted in different color channels. Furthermore, it uses a generalization of the Hough Transform introduced by Loy and Barnes [6]. This is motivated by the fact that the targeted objects are all instances of equiangular polygons, including circles which can be thought of as equiangular polygons with infinite number of sides.

Original regular polygon transform is augmented with an appropriate image preprocessing intended to locate the regions of interest rapidly and further enhance the edges of specific color within each. The former goal is achieved using an attentive cascade of classifiers introduced in [14]. Each weak classifier is a 1D perceptron associated with one of the Haar wavelet features shown in Fig. 2. Similarly to [1], we additionally parametrize these features with color. Three possible assignments are: red, blue and yellow which correspond to the most characteristic colors of the road signs. To be able to evaluate such features, we filter an input image by enhancing the sign-specific colors. This enhancement for each RGB pixel $\mathbf{x} = [x_R, x_G, x_B]$ and $s = x_R + x_G + x_B$ is achieved using the following set of transformations:

$$\begin{array}{l} f_R(\mathbf{x}) = \max(0, \min(x_R - x_G, x_R - x_B)/s) \\ f_B(\mathbf{x}) = \max(0, \min(x_B - x_R, x_B - x_G)/s) \\ f_Y(\mathbf{x}) = \max(0, \min(x_R - x_B, x_G - x_B)/s) \end{array} . \qquad (4)$$

For each color-enhanced image the corresponding integral image required by the Haar features is then calculated [14].

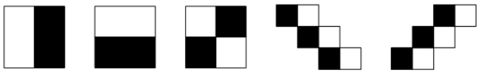

Fig. 2. Haar wavelet features used for sign detection.

Cascade is trained separately for different scales on the dataset comprised of several hundred pre-labeled positive images and several thousand randomly generated negative images, both taken from the traffic video sequences. The cascade parameters are chosen in such a way that practically all true signs are captured while the false positive rate is below 1%. In system runtime, the redundant hypotheses generated by our attentive detector around the true signs are clustered and a single accurate bounding box per candidate is estimated in a similar way as is proposed in [13].

In each region containing the sign candidate, color-specific edge maps are extracted by a simple filter which for a given pixel picks the highest difference among the pairs of neighboring pixels that could be used to form a straight line through the middle pixel being tested. Obtained values are further thresholded and only in the resulting edge pixels values of directional and magnitude gradient are recorded. This method is adequate to our problem as it enables a quick extraction of edges and avoids expensive computation of the whole gradient magnitude map which, with the exception of the sparse edge pixels, is of no use to the shape detector. Extracted color gradient maps provide the sufficient data for initialization of the Pixel Relevance Model. Specifically, we consider the measured gradient magnitude at pixel x_{ij} to be an observable symptom of the unknown pixel relevance and use it to initialize the first estimate of the state, i.e. $r_{ij}(0) = g_{ij}$.

For a given pair of edge and gradient images associated with a given interest region and the feature color c, the appropriate instance of the Loy and Barnes's shape detector [6] is run to yield a sign candidate of the known scale. For instance, for a "yellow pair" a triangular shape detector is triggered to search for the yellow cautionary sign. As each found candidate has known shape and border color c, detector serves as a pre-classifier reducing the number of possible templates to analyze in the later stage to the ones contained in either category. The major modification of the original equiangular polygon detector we allow is the following. Although it still uses the directional gradient information to converge the votes coming from different sides of a polygon in its alleged centroid [6], strength of these votes is determined by the pixel relevance, instead of magnitude of the gradient.

C. Tracking

Once a candidate sign is detected, it is unnecessary to search for it in the consecutive frames in every possible location. Our road sign tracker is composed of three complementary parts: 1) Kalman Filter [5] responsible for maintaining a localized search region around the expected position of the candidate, 2) the Pixel Relevance Model, discussed in section II A, used to update a belief on the relevance of the pixels contained in this region, and 3) a regular shape detector [6]. The latter is run on the current color-specific edge, directional gradient, and posterior pixel relevance maps extracted within the interest region to yield the best-matching contour of a sign being tracked.

The Kalman Filter tracks the signs in a geometric fashion. The state of the tracker is given by two vectors: $\mathbf{S}_c = [x, y, v_x, v_y]^T$, and $\mathbf{S}_s = [w, h]^T$. $\mathbf{S}_c$ encodes the coordinates of the sign's centroid and its momentary velocity, $\mathbf{v} = [v_x, v_y]$, in the 2D image plane. The size of a sign expressed in terms of width and height of its minimum bounding rectangle is given by $\mathbf{S}_s$. Position and size of the current-frame search region is determined from the prior estimate of the filter parameters and their variances. Velocity $\mathbf{v}$, used in the prediction equation of the filter, is estimated from the point correspondences established in the consecutive frames in the interior of the sign as last detected. Correction step is performed upon capturing the candidate shape at time $t+1$ by the Loy and Barnes' detector [6].

Our Kalman tracker has another interesting feature. The prior search region estimate can be used to modulate the relevance of the contained pixels by highlighting these that lie on the hypothesized motion-compensated contour of a sign or in its vicinity, and diminishing the importance of the remaining ones. As a result, the unwanted edges inside and around the sign being tracked become suppressed and the shape detector can produce more accurate fits. Knowledge of the previous Kalman Filter state estimate and the current-frame velocity measurement provides the necessary data for the abovementioned motion compensation. The influence of the filter on the actual pixel relevance map maintained in the spatio-temporal graph introduced in section II A is incorporated in the prediction process (1). More specifically, we make the transition function $\phi_{ij}(t)$ dependent on the current variance of the state parameters. A hypothetical contour of a sign together with the search region are translated from its centroid's estimate at time t to the new, prior estimate at time $t+1$. Then, a local Distance Transform (DT) [2] is computed in such a motion-compensated region. Finally, the transition function of the Pixel Relevance Model is computed as:

$$\phi_{ij}(t+1) = e^{-k_D d_{DT}(i,j)} \left(1 - e^{-k_T r_{N(i,j)}(t)}\right) \quad , \qquad (5)$$

where $d_{DT}(i,j)$ denotes the appropriate distance value picked from the DT image, $k_D \sim \frac{1}{E^2}$ and E is the average of the prior variances of the Kalman Filter state parameters at time $t+1$. The effect of Distance Transform on the pixel relevance map is visualized in Fig. 3.

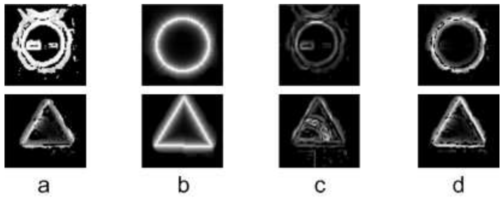

Fig. 3. Different stages of pixel relevance processing in a single video frame around a tracked candidate road sign: a) relevance map at time t, b) prior relevance map at time $t+1$ (after KF-regularized projection), c) gradient magnitude map at time $t+1$ (observable evidence), d) posterior relevance map (after incorporating observation). In all images intensity is scaled to the range $[0, 1]$ for better visualization.

The circle/regular polygon detector is integrated with the above tracking framework as a third component to produce the final shape which is then directly processed by the classifier discussed in the next section. The detector is run on the edge maps and the directional gradient maps of the current frame but uses an *a posteriori* map of pixel relevance to determine the strength of the votes.

To summarize, the main role of the Kalman Filter in the above framework is to maintain a reasonably small region of interest for the pixel relevance evolution model. However, an additional task of the filter is to adjust the parameters of this model on-the-fly. As a result, for very accurately estimated search region the pixel relevance is peaked at the expected contour of a sign, as in Fig. 3d, and the unwanted edges that with high probability are part of the sign's interior or cluttered background, become suppressed. This in turn directs the focus of the shape detector to the sign's centroid. However, when the uncertainty on the Kalman Filter state is high, our confidence of having captured the sign area well is lower accordingly. In this case the modulation factor is relaxed and the relevance model is to a larger extent allowed to "evolve on its own". Consequently, the relevance image is much more blurry and better reflects the full map of the currently observed gradients, giving the shape detector an opportunity to recover from a likely poor fit.

III. SIGN RECOGNITION

Once a candidate sign has been detected, it has to be identified. As stated in section II B, a general category of a sign is determined by its shape and the color of its rim at the time when it is for the first time detected. Nevertheless, there are still up to 55 signs per category, some being very similar to one another. This makes the recognition a challenging task. The sign discrimination method adopted in this work is based on an idea of building a compact, class-specific representation of each sign by picking these fragments of the pictogram in which it differs the most from all other signs in the same category. Furthermore, as the model signs are well-defined, we think such representation can be inferred without recourse to the real-life images. Below, an outline of our discriminative sign representation and the classifier design is given. Details can be found in [12].

A. Discriminative Local Region Representation

To represent the pictograms of signs, we consider a space of all local, non-overlapping square regions of the template sign images. The goal is to select the most discriminative regions out of this space so that the unseen signs are the most easily distinguishable for the classifier. Formally, assuming a pre-determined category of signs $C = \{T_i : i = 1, \ldots, N\}$ and a candidate image x_j, our goal is to determine the class of x_j by maximizing posterior:

$$p(T_i|x_j,\theta_i) = \frac{p(x_j|T_i,\theta_i)p(T_i)}{\sum_{k=1}^{N} p(x_j|T_k,\theta_k)p(T_k)} \ . \tag{6}$$

Varying indices of the model parameter vector θ are used to emphasize the fact of allowing different model parameters $\theta_i = (\mathbf{I}_i, \mathbf{W}_i)$ for each template T_i. The first parameter defines a binary vector determining which local regions of the image to incorporate in the target region set $\mathbf{S}_i$. The second parameter defines a vector of corresponding region weights. The optimal parameters of each model, θ_i^*, are learned through maximization of the objective function:

$$O(\theta_i) = \sum_{j \neq i} \widehat{d}_{\mathbf{S}_i}(T_j, T_i) \ , \tag{7}$$

where the dissimilarity $\widehat{d}_{\mathbf{S}_i}(T_j, T_i)$ is defined as a mean of the dissimilarities between the images I and J measured within the individual regions r_k contained in the set $\mathbf{S}_i$:

$$\widehat{d}_{\mathbf{S}_i}(I, J) = \frac{1}{|\mathbf{S}_i|} \sum_{k=1}^{|\mathbf{S}_i|} d_{r_k}(I, J) \ . \tag{8}$$

Dissimilarity between two images within a single region r_k is based on the notion of Distance Transform (DT) [2]. Specifically, for each distinct color in the template sign image a separate DT is calculated as though that image was a binary map with the pixels of that color being the feature pixels, and all other pixels being the non-feature pixels. We refer to such collection of color-specific DT images as a Color Distance Transform (CDT). The CDT images computed for a sample road sign template are shown in Fig. 4.

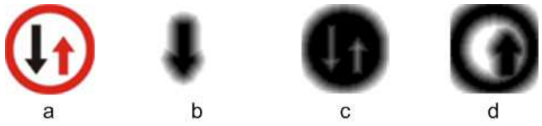

Fig. 4. Color Distance Transform (CDT) images: a) original discretized color image, b) black DT, c) white DT, d) red DT. Darker regions denote shorter distance. This figure is best viewed in color.

With CDT available, $d_{r_k}(I, J)$ can be expressed as:

$$d_{r_k}(I, J) = \frac{1}{m^2} \sum_{x,y} \widetilde{d}_{CDT}(I(x,y), J(x,y)) \ , \tag{9}$$

where for each pixel with coordinates (x, y) contained in the $m \times m$ pixel region, distance $\widetilde{d}_{CDT}(I(x,y), J(x,y))$ is picked from the appropriate CDT image of J, depending on the discrete color of this pixel in I.

Maximization of the objective function in (7) is implemented as a greedy forward selection algorithm [11] which operates solely on the template sign images. Details of this procedure can be found in [12]. On output it yields for each template a variable-size set of its most unique regions, i.e. those in which this template differs the most from all other same-category templates. Additionally, the weight associated with each region reflects its individual discriminative power. The key feature of our training algorithm is that it captures the same amount of dissimilarity in each class-specific model, regardless of the actual number of regions extracted. This amount can be controlled with a global parameter, t_d, called *dissimilarity threshold*. This feature makes the signs directly comparable and hence facilitates classification. Visualization of the most discriminative regions for several cautionary signs is given in Fig. 5.

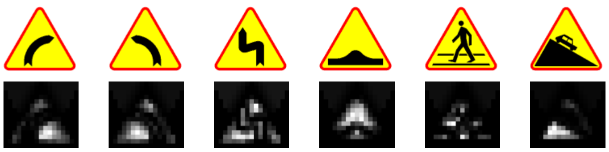

Fig. 5. A sample of triangular cautionary signs (above) and the discriminative local regions obtained for parameter $t_d = 0.7$ (below). Brighter regions correspond to higher dissimilarity.

B. Dynamic Recognition from Video Input

A detected sign candidate is subject to preprocessing before it can be classified. First, its orientation is corrected based on the information provided by the detector [6]. In the next step the region corresponding to the minimum bounding rectangle of the sign is cut out of the image and scaled to a common size, typically 60×60 pixels. The possible, shape-dependent background fragments are masked out. In the resulting image the color space is collapsed to a few basic colors using a Gaussian Mixture color classifier trained on the real-life road sign images [12].

The normalized image of a sign is classified in each frame of the input video according to the Maximum Likelihood approach by template matching. However, matching to each template is done with respect to the class-specific representation determined in the training process. Assumptions of 1) Gaussian distribution of the local region dissimilarities and 2) equal class priors $p(T_i)$ make it possible to convert the maximization of likelihood to the minimization of distance. Therefore, according to (6), a winning class $L(x_t)$ of the unknown observed candidate x_t at time t is determined by:

$$\begin{aligned} L(x_t) = \;& \arg\max_i p(x_t|T_i, \theta_i) = \\ & \arg\min_i \widehat{d}_{\mathbf{S}_i,\mathbf{W}_i}(x_t, T_i) \end{aligned} \quad , \tag{10}$$

where the regions in $\mathbf{S}_i$ and the corresponding weights in $\mathbf{W}_i$ denote the ones learned in the training stage for the template T_i. Distance $\widehat{d}_{\mathbf{S}_i,\mathbf{W}_i}(x_t, T_i)$ is defined as:

$$\widehat{d}_{\mathbf{S}_i,\mathbf{W}_i}(I, J) = \frac{\sum_{k=1}^{|\mathbf{S}_i|} w_k d_{r_k}(I, J)}{\sum_{k=1}^{|\mathbf{S}_i|} w_k} \quad . \tag{11}$$

The above metric is a weighted version of the mean dissimilarity defined in (8). It requires an on-line discretization of the observed sign image (possible to be implemented efficiently using color lookup table) and offline-computed Color Distance Transforms of the templates.

Classification results for the individual frames are integrated through the whole sequence over time. Hence, at a given time point t all the observations made since the sign was for the first time detected until t are incorporated in the current classifier's decision. Assuming independence of the observations from the consecutive frames, this decision is determined by picking the smallest cumulative distance from the template:

$$L(X_t) = \arg\min_i \sum_{k=1}^{t} q(t) \widehat{d}_{\mathbf{S}_i,\mathbf{W}_i}(x_k, T_i) \quad , \tag{12}$$

where $q(t) = b^{t_{last} - t}$, $b \in (0, 1]$, is a relevance of the observation x_t. As t_{last} means the time point when the sign is for the last time seen, the observation's relevance is made dependent on the candidate's age (and thus size) to reflect an empirical fact that an object becomes clearer while being approached by the camera.

IV. EXPERIMENTAL RESULTS

To evaluate the proposed road sign recognition approach, 96 video sequences containing 164 signs in total (out of which 48 unique) were recorded during daytime with a standard DV camcorder mounted in front of the car's windscreen. This data were collected in a variety of urban, countryside and freeway scenes. Overall recognition results are shown in Tab. 1. It is assumed that for a given test sequence the ultimate classifier's decision is the one made at the time when the track of the sign is permanently lost as a result of it getting out of the camera's field of view.

	t_d	RC (55)	BC (25)	YT (42)	BS (13)	All (135)
detected	–	85.2%	100.0%	96.8%	87.8%	93.3%
recognized	0.97	95.7%	91.2%	83.3%	88.9%	88.2%
	0.9	95.7%	94.1%	86.7%	97.2%	92.2%
	0.7	95.7%	100.0%	85.0%	86.1%	90.2%
	0.5	95.7%	100.0%	81.7%	83.3%	88.2%
	best	95.7%	100.0%	86.7%	97.2%	93.5%

TABLE I

RECOGNITION PERFORMANCE FOR DIFFERENT VALUES OF DISSIMILARITY THRESHOLD t_d AND TEMPORAL WEIGHT BASE $b = 0.8$. THE NUMBER OF CLASSES IN EACH CATEGORY: RED CIRCLES (RC), BLUE CIRCLES (BC), YELLOW TRIANGLES (YT), AND BLUE SQUARES (BS) IS GIVEN IN PARENTHESES.

The obtained detection results, as seen in Tab. 1, depend on the sign category. However, the figures may be biased by the fact that different categories were not equally represented in the test data. An overall detection rate of over 93% is good for the data we collected and the large size of our template database. The average processing speed of 20-25 fps was achieved and only a few false alarms across all test sequences were reported. All true positives were detected and tracked at a physical distance from the camera of approximately 10-30 meters. The majority of detection failures were caused by the insufficient contrast between a sign's boundary and the background, especially for the signs appearing in shade or seen against sunlight. In a few cases this low contrast was caused by the faded dye on the sign's plate.

Figure 6 shows examples of the two sequences where a sign is being tracked over time. Each upper row of images illustrates the actual gradient magnitude map whereas the images in each lower row depict the corresponding posterior pixel relevance maps. It can be noticed that in the pixel relevance images the sign contours are clearly emphasized but the other high-gradient image regions tend to be suppressed. Thanks to the accurate motion compensation of the pixel relevance map provided by the Kalman Filter we found this technique very useful in combination with the shape detector of Loy and Barnes [6]. As the latter relies on the contour of the shape being searched for, possibility of the inaccurate detection in each frame is greatly reduced.

For the best set of dissimilarity thresholds we have managed to reach over 93% correct classification rate, i.e. the percentage of the correctly classified signs among these that were detected. This figure makes our method comparable to the recently published ones [1], [8]. However, it should be noted that our template database contains 135 model signs – significantly more than in any of the previous studies. Direct comparison with the respective algorithms is not possible as neither the test data nor the details of data acquisition are disclosed. The optimal dissimilarity threshold values were determined experimentally. We observed that for each sign category this threshold must strike a balance between maximizing template signs' separability and the reliability of the obtained dissimilarity information in the real-data context. Very high threshold values lead to the separation of very few good regions for a particular model sign. However, for a noisy video frame image this sparse information is usually insufficient as the relevant fragments of the pictogram may look distorted, blurred, or their color may vary. We found a sensible limit for the dissimilarity threshold to be 0.98–0.99 as for higher values of t_d representations of certain signs become empty. Very low threshold values on the other hand introduce information redundancy by allowing image regions that contribute little to the uniqueness of a given sign. In a resulting feature space these signs look more similar to one another and are hence more difficult to tell apart.

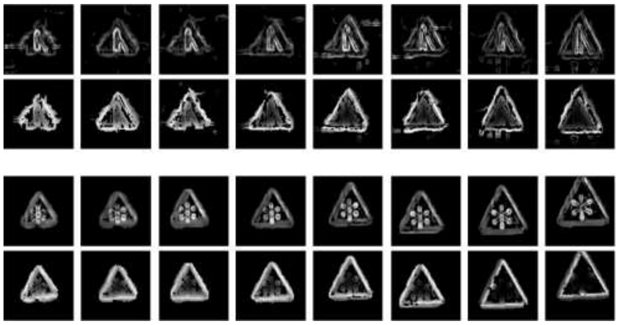

Fig. 6. Sample sequences of a road sign being tracked from the perspective of gradient magnitude (upper rows) and pixel relevance (lower rows). While the gradient maps contain all high-magnitude regions, in the relevance images the non-contour peaks of the gradient tend to be suppressed. Each image corresponds to the local search region at a given time point and for illustration purpose these images are scaled to equal size.

The main cause of the observed classification errors is the sensitivity of our template matching method to the accuracy of the detector. This effect is even strengthened by the fact that a large gamut of signs is focused on. Many of these errors were a result of confusion between the nearly identical classes, especially in the category of triangular cautionary signs. In these cases the correct template frequently received the second best score. In only several sequences classification failed due to the inaccurate color segmentation. This proves usefulness of the Gaussian Mixture color modeling. It should be noted that a significant gain in the classification accuracy was achieved owing to the temporal integration scheme we used to combine the individual frame observations. A tendency of the most recent observations being the most useful was clearly confirmed by the experiments.

V. CONCLUSIONS

In this paper we have presented a comprehensive approach to traffic sign recognition from video input. Our detection module consists of four components: (1) an attentive classifier cascade used to quickly discard the irrelevant fragments of the scene, (2) an equiangular polygon detector used in the remaining fragments to capture the regular contour of a sign, be it a circle, triangle or square, (3) a Kalman Filter to reduce the search region around the candidate signs already being tracked, and (4) the Pixel Relevance Model to provide a confidence measure for a pixel on the sign's contour. The pixel relevance is computed initially from the previous detection, and updated dynamically from the current observation of gradient magnitudes. It helps emphasize the contour of a sign and suppress the irrelevant information from the background and inside the sign, which in turn improves the accuracy of the circle/polygon detector. At the recognition stage, a discriminative local region representation of signs is proposed. It is constructed directly from the template sign images so as to capture possibly the most significant differences between them. It has been shown that the obtained discriminative local image regions can be used in conjunction with a conventional classifier operating by class-specific, temporally integrated template matching. Additionally, the smooth distance metric provided by the Color Distance Transform (CDT) makes this matching resistant to minor misalignments introduced by the shape detector.

We have evaluated the proposed approach on real traffic videos. Overall, a recognition rate of over 93% has been achieved with a processing speed of 20-25 fps and a decently low number of false positives. It is also important to note that the size of the template database used in our experiments is significantly greater than those reported previously.

REFERENCES

[1] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, and T. Koehler. A system for traffic sign detection, tracking and recognition using color, shape, and motion information. In *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 255–260, 2005.

[2] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34(3):344–371, 1986.

[3] A. de la Escalera, L. E. Moreno, M. A. Salichs, and J. M. Armingol. Road traffic sign detection and classification. *IEEE Trans. on Industrial Electronics*, 44(6):848–859, 1997.

[4] C.-Y. Fang, S.-W. Chen, and C.-S. Fuh. Road-sign detection and tracking. *IEEE Trans. on Vehicular Technology*, 52(5):1329–1341, 2003.

[5] R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. of the ASME - Journal of Basic Engineering*, 82:35–45, 1960.

[6] G. Loy, N. Barnes, D. Shaw, and A. Robles-Kelly. Regular polygon detection. In *Proc. of the 10th IEEE Int. Conf. on Computer Vision*, volume 1, pages 778–785, 2005.

[7] J. Miura, T. Kanda, and Y. Shirai. An active vision system for real-time traffic sign recognition. In *Proc. of the IEEE Conf. on Intelligent Transportation Systems*, pages 52–57, 2000.

[8] P. Paclík, J. Novovicová, and R. P. W. Duin. Building road-sign classifiers using a trainable similarity measure. *IEEE Trans. on Intelligent Transportation Systems*, 7(3):309–321, 2006.

[9] P. Paclík, J. Novovicova, P. Pudil, and P. Somol. Road sign classification using the laplace kernel classifer. *Pattern Recognition Letters*, 21(13–14):1165–1173, 2000.

[10] G. Piccioli, E. D. Micheli, P. Parodi, and M. Campani. A robust method for road sign detection and recognition. *Image and Vision Computing*, 14(3):209–223, 1996.

[11] P. Pudil, J. Novoviová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[12] A. Ruta, Y. Li, and X. Liu. Towards real-time traffic sign recognition by class-specific discriminative features. In *Proc. of the 18th British Machine Vision Conference*, volume 1, pages 399–408, 2007.

[13] A. Ruta, Y. Li, and X. Liu. On-line human detection using an attentive classifier cascade. Technical report, Brunel University, 2008.

[14] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.