



Real-time traffic sign recognition from video by class-specific discriminative features

Andrzej Ruta*, Yongmin Li, Xiaohui Liu

School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

ARTICLE INFO

Article history:

Received 1 August 2007

Received in revised form 22 May 2009

Accepted 26 May 2009

Keywords:

Traffic sign recognition

Computer vision-based driver assistance

Discriminative local regions

Colour Distance Transform

Forward feature selection

ABSTRACT

In this paper we address the problem of traffic sign recognition. Novel image representation and discriminative feature selection algorithms are utilised in a traditional three-stage framework involving detection, tracking and recognition. The detector captures instances of equiangular polygons in the scene which is first appropriately filtered to extract the relevant colour information and establish the regions of interest. The tracker predicts the position and the scale of the detected sign candidate over time to reduce computation. The classifier compares a discrete-colour image of the observed sign with the model images with respect to the class-specific sets of discriminative local regions. They are learned off-line from the idealised template sign images, in accordance with the principle of one-vs-all dissimilarity maximisation. This dissimilarity is defined based on the so-called *Colour Distance Transform* which enables robust discrete-colour image comparisons. It is shown that compared to the well-established feature selection techniques, such as Principal Component Analysis or AdaBoost, our approach offers a more adequate description of signs and involves effortless training. Upon this description we have managed to build an efficient road sign recognition system which, based on a conventional nearest neighbour classifier and a simple temporal integration scheme, demonstrates a competitive performance in the experiments involving real traffic video.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Recognition of traffic signs has been a challenging problem for many years and is an important task for the intelligent vehicles. Although the first work in this area can be traced back to the late 1960s, significant advances were made later, in the 1980s and 1990s, when the idea of computer vision-based driver assistance attracted worldwide attention and the video processing became more attainable. Originating from the large-scale projects developed in the USA, Europe and Japan, intensive research on traffic sign recognition is nowadays conducted by both academic and industrial groups all over the world, the latter often being in strong relation to the car industry. Despite all this effort being made and the driver's comfort and safety being at stake, surprisingly few working recognition systems of this kind are at present in operation. It indicates that the human driver still remains the best guarantor of safety in the traffic environment.

Road signs have unique properties distinguishing them from the multitude of other outdoor objects. These properties were exploited in numerous approaches to the detection and recognition of signs. In a majority of published work a two-stage sequential approach was adopted, aiming at locating the regions of interest and verifying the

hypotheses on the sign's presence (detection), and subsequently determining the type of the detected sign (recognition) [7,9,13,25]. To detect possible sign candidates, traditionally colour information was primarily used [7,9,25], followed by the geometrical edge analysis [7,23,25,26] or corner analysis [9]. Based on the idealised templates of road signs, Fang et al. [17] built two separate neural networks to extract relevant colour and shape features of signs which were further integrated in a fuzzy way. This approach was reported to be accurate but computationally very intensive.

Some authors preferred a strictly colourless approach as they did not consider the colour segmentation absolutely reliable due to its sensitivity to various factors, such as distance from the target, weather conditions, time of day, or reflectance of the signs' surfaces. These approaches for example utilised genetic algorithms [8] or distance transforms (DTs) [10]. In [14], where the colour information was also not considered, the images were transformed using wavelets and classified by a Perceptron neural network. Not exploiting the colour information does not preclude successful discrimination between certain types of signs. However, in presence of similar pictograms from different semantic categories, and when there are a large number of classes in the database, colour carries priceless discriminative information that in our opinion should be used whenever possible. Ultimately, in certain countries, e.g. Japan, there are pairs of different signs in the highway code that, when converted to grey scale, look exactly the same. To tell them apart, colour information is absolutely necessary.

* Corresponding author.

E-mail addresses: Andrzej.Ruta@brunel.ac.uk, mailme@aruta.pl (A. Ruta), Yongmin.Li@brunel.ac.uk (Y. Li), Xiaohui.Liu@brunel.ac.uk (X. Liu).

Among the studies in which the colour information was used to detect the traffic signs, majority based on the non-RGB colour spaces. Hue-Saturation-Value (HSV) colour model was the most popular one as being based on human colour perception. Additionally, it is considered largely invariant to illumination changes. Piccioli et al. [7] determined the regions of interest by clustering small blocks of the image where the number of pixels having value of hue in an appropriate range exceeded a predefined threshold. HSV model was also used in [13] to classify the test sign images into several distinctive categories and in [17] to extract colour features by a neural network. Other colour appearance models were less popular. For instance, Escalera et al. [9] preferred to operate on the ratios between the intensity of a given channel and the sum of all RGB channel intensities. They pointed out that the RGB-HSV conversion formulas are non-linear and hence the computational cost involved is too high. However, this problem can be easily avoided by pre-computing the colour space conversion and storing it in a look-up table.

The CIECAM97 colour model was chosen by Gao et al. [25] for image segmentation. The authors found it suitable to estimate the appearance of sign-characteristic colours independently from real-life images taken under different viewing conditions. An interesting and robust approach to colour-based detection and segmentation of road signs using IHLS colour space was also proposed in [20]. In both above studies the device-independent colour appearance models were adopted, which is a reasonable choice compared to the standard device-dependent RGB model. Bahlmann et al. [24] used a completely different strategy. They trained the scale-specific road sign detectors using Haar wavelet features [21] parametrised by colour. The best features were selected within a cascaded AdaBoost framework from a large space of features defined over multiple colour representations: plain R, G, and B channels, normalised R, G, and B channels, and a grey-scale channel. Therefore, the most suitable colour representation was inferred automatically from the data rather than arbitrarily chosen by a human. The main disadvantage of this method is its demand for large amounts of training images and generally strenuous training. Besides, for a 384×288 video resolution only 10 fps processing speed was achieved. Finally, we think that the exceptional discriminative power of the detector presented by Bahlmann and his colleagues was only possible to achieve because a very narrow category of signs exhibiting similar appearance characteristics was considered in their work.

In several studies the problem of tracking of the existing sign candidates was given consideration [7,15,17,26]. However, reliable prediction of the geometrical properties of signs from a moving vehicle is complex in general as the vehicle's manoeuvres are enforced by the actual traffic situation and therefore cannot be *a priori* known. To overcome this problem, the above approaches imposed simplified motion model, e.g. assuming constant velocity. The work of Miura et al. [15] deserves a particular attention as the authors proposed a two-camera active vision system for capturing and tracking the road sign candidates. It was composed of a wide-angle camera that detected the candidates within the entire scene, and a telephoto camera which was directed to the predicted position of each candidate to capture it in a larger size in order to extract the necessary pictogram details. The apparent limitation of the system introduced by Miura et al. is its dependency on the hardware. In particular, the telephoto camera requires substantial amount of time to change the viewing direction. Besides, the system seems to be able to recognise only one sign at a time.

The work of Escalera et al. [22] is in the minority of studies where modelling of the full structure of the apparent affine motion of a sign in the image plane was attempted. In this approach a deformable model of sign was developed which enabled detection of geometrically distorted, poorly illuminated, noised and occluded targets. The state of a detected sign's geometry was updated based on

the previous-frame state according to an affine motion matrix with found deformation parameters. Two search strategies for obtaining the optimal values of these parameters were proposed: genetic algorithm and simulated annealing. This indeterministic search triggered in every frame of the input video was driven by minimisation of a sum of relatively complex energy functions relating the deformation parameters to the information about the colour and the shape of the observed traffic sign. Promising results were reported using this method, but it is far too slow for real-time execution in the realistic driver support systems.

At the classification stage a pixel-based approach was most often adopted and the class of the detected sign was determined by the cross-correlation template matching [7,22] or neural network [9]. Feature-based approach was also frequently used. For instance in [13] authors utilised various statistical characteristics, e.g. moments, calculated from the binary images of the inner parts of the detected road sign candidates. Gao et al. [25] classified the signs by comparing the 49-dimensional feature vectors encoding their local edge orientations and density at the arbitrary fixation points to the corresponding vectors computed for the template sign images. Up to 95% success recognition rate was achieved in the experiments involving still camera images. Manual feature selection is, however, unconvincing.

In the abovementioned work of Bahlmann et al. [24] a Bayes classifier was used to fuse the individual observations over time, assuming Gaussian distribution of the feature vector and the independence of consecutive frame observations. Only 6% error rate and 10 fps average processing speed was reported using this method on the 30-min test video. However, only a narrow subset of signs were targeted in this work—speed limit signs and “no passing” signs. Paclík et al. [27] introduced a different strategy built upon the idea that a candidate sign can be represented as a set of similarities to the stored prototype images. For each class similarity assessment was made with respect to a different set of local regions refined in the training process. For relatively simple problems involving significantly dissimilar signs this approach offers competitive performance.

A final remark concerning the state-of-the-art classification of road signs should be made to emphasise that in the majority of previous studies only the narrow subsets of signs or traffic situations were considered. Typically, authors only focused on a single semantic category, e.g. the speed limit signs, or the relatively dissimilar signs from multiple categories, which facilitated the recognition enormously. In many studies the problem was even more simplified by restricting it to the recognition of signs on the static, sometimes already pre-segmented images. The experiments on the relatively large sign databases were conducted in: [13] (50 signs from multiple categories, but static images only), [22] (83 signs from multiple categories, but essentially no video sequences), [25] (87 signs from three categories, but artificially generated and noised).

In this work we have developed a two-stage symbolic road sign detection and classification system. Fig. 1 shows a screenshot illustrating how this system detects and recognises a sign in a sample frame of the input video. More specifically, our detector is a form of a well-constrained circle/regular polygon detector, similar to the one used by Loy et al. [23] and augmented with the appropriate colour pre-filtering. The Kalman filter (KF)-based tracker is additionally employed in each frame of the input video to predict the position and the scale of a previously detected candidate and hence to reduce computation. In the classification stage, motivated by [27], we introduce a novel feature selection algorithm that extracts for each sign a small number of critical local image regions encompassing the most dissimilarity between this sign and all other signs. Within these regions, robust image comparisons are made using a distance metric based on what we call *Colour Distance Transform* (CDT), which enables efficient pictogram classification.

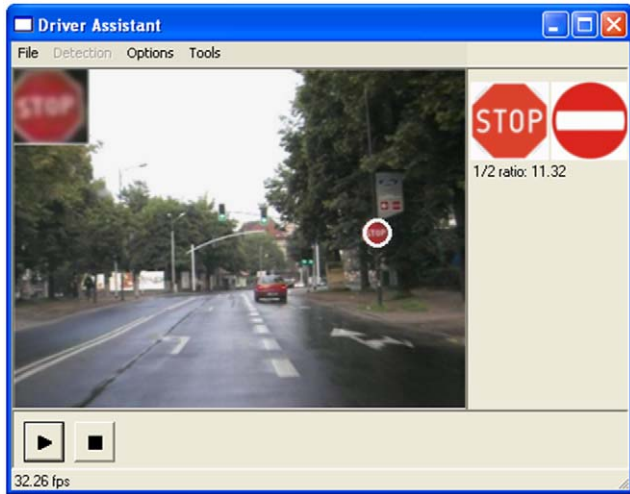


Fig. 1. Screenshot from our traffic sign recognition system in action. The best-scored and the second best-scored template are shown.

This paper is an extension to our previous work [28,29] in that it contains a much more detailed description of the ideas and algorithms as well as an experimental justification of the design choices made in the aforementioned studies. In particular, a more thorough analysis of the discriminative representation of traffic signs is provided and the strengths of the discrete-colour image representation and the *Colour Distance Transform* are shown experimentally. Moreover, the impact of the class-specific feature space's dimensionality on the overall recognition accuracy is shown and the proposed classifier's error rate is compared to the error rates of the classifiers learned via PCA and AdaBoost using a large traffic sign image dataset.

The rest of this paper is organised as follows: In Section 2 traffic sign detection and tracking are outlined. Sections 3 and 4 discuss the main contributions of this work, discrete-colour/CDT image representation and a discriminative local feature selection algorithm. In Section 5 a temporal sign classification framework is described. Section 6 contains the experimental evaluation of our approach, involving both: static sign images and real traffic video. Finally, conclusions are drawn in Section 7.

2. Sign detection and tracking

Our road sign detector is triggered every several frames for a range of scales to capture new candidates emerging in the scene. Because these initial detections take place relatively early, when the signs undergo nearly no motion and do not grow in the image plane, the risk of missing a sign is minimal. However, the computational effort is significantly reduced. The detector encompasses *a priori* knowledge about the model signs, uniquely identified by their general shape, characteristic colours, and pictogram. Based on the first two properties, four sign categories coinciding with the well-known semantic families are identified: instruction signs (blue circular), prohibitive signs (red circular), cautionary sign (yellow triangular), and informative signs (blue square).¹

As we believe the shape and the rim colour are sufficient visual cues to locate the signs reliably, the proposed detector operates on the colour gradient and edge maps of the original video frames. Furthermore, it uses the approach of Loy et al. [23], in which a recipe is given for how to locate the most likely instances of

equiangular polygons in the image, which is considered to be drawn from a mixture of regular polygons. Note that the non-circular road signs considered in this study are all instances of such polygons. The realisation of the algorithm of Loy and his colleagues very much resembles a circular Hough Transform (HT), where shape localisation is based on voting in the parameter space. Because circles can be thought of as regular polygons with the infinite number of sides, this method can be treated as a generalisation of HT.

Original regular polygon transform is augmented with the appropriate image preprocessing intended to locate the regions of interest (Rols) and further enhance the edges of specific colour within each. For the first task a traditional approach known from many previous works on road sign recognition is adopted. First, the whole scene image is colour-discretised using a reasonable set of fixed thresholds in a Hue-Saturation-Value colour space, similar to this used in [12]. In the resulting image every pixel is assigned one of the six colours: black, white, red, yellow, green, or blue. The image is then divided into small 20×20 -pixels blocks. Now, for each of the four distinguished sign categories blocks are independently marked as either "feature" or "non-feature", depending on the percentage of the contained pixels having the colour characteristic to the respective category, i.e. red for the prohibitive signs, blue for the instruction and informative signs, and yellow for the cautionary signs. As a result, for each of the three relevant colours a separate binary map of feature blocks is obtained. Subsequently, a region growing algorithm is fed by each of the feature block maps to form blobs containing connected blocks of respective colours. The final rectangular interest regions are constructed by picking the bounding rectangle of each blob. An additional margin of width equal to half of the radius of the largest considered signs is added on each side of each Rol to avoid capturing incomplete signs that appear partially occluded or indistinct. From now on, further processing is localised in the found Rols, instead of the whole scene.

In the extracted Rols edges and gradients of the respective colours need to be extracted for the shape detector to work. Therefore, in each interest region the pixels are first transformed so that the colour associated with this Rol is enhanced. For each RGB pixel $\mathbf{x}=[x_R, x_G, x_B]$ and $s = x_R + x_G + x_B$ a simple colour enhancement is provided by a set of transformations:

$$f_R(\mathbf{x}) = \max(0, \min(x_R - x_G, x_R - x_B)/s),$$

$$f_B(\mathbf{x}) = \max(0, \min(x_B - x_R, x_B - x_G)/s),$$

$$f_Y(\mathbf{x}) = \max(0, \min(x_R - x_B, x_G - x_B)/s). \quad (1)$$

Transforms defined in (1) effectively extract the red, blue, and yellow image fragments. First two extract these parts of the image where the red or blue component, respectively, dominates the most over both remaining components. The third formula has similar meaning, but as the pure yellow colour has equal value in the red and green channels and zero in the blue channel, it attempts to enhance pixels where both former components dominate the most over the latter. Examples illustrating the effect of filters (1) applied to the original RGB traffic images are given in Fig. 2. In the resulting images colour-specific edge maps are extracted by a simple filter which for a given pixel picks the highest difference among the pairs of neighbouring 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 calculated. This technique 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.

For a given pair of gradient and edge images associated with colour c , the appropriate instances of the regular polygon detector [23] are run to yield a set of possible sign shapes in the predefined

¹ In many countries the cautionary signs have white rather than yellow background.

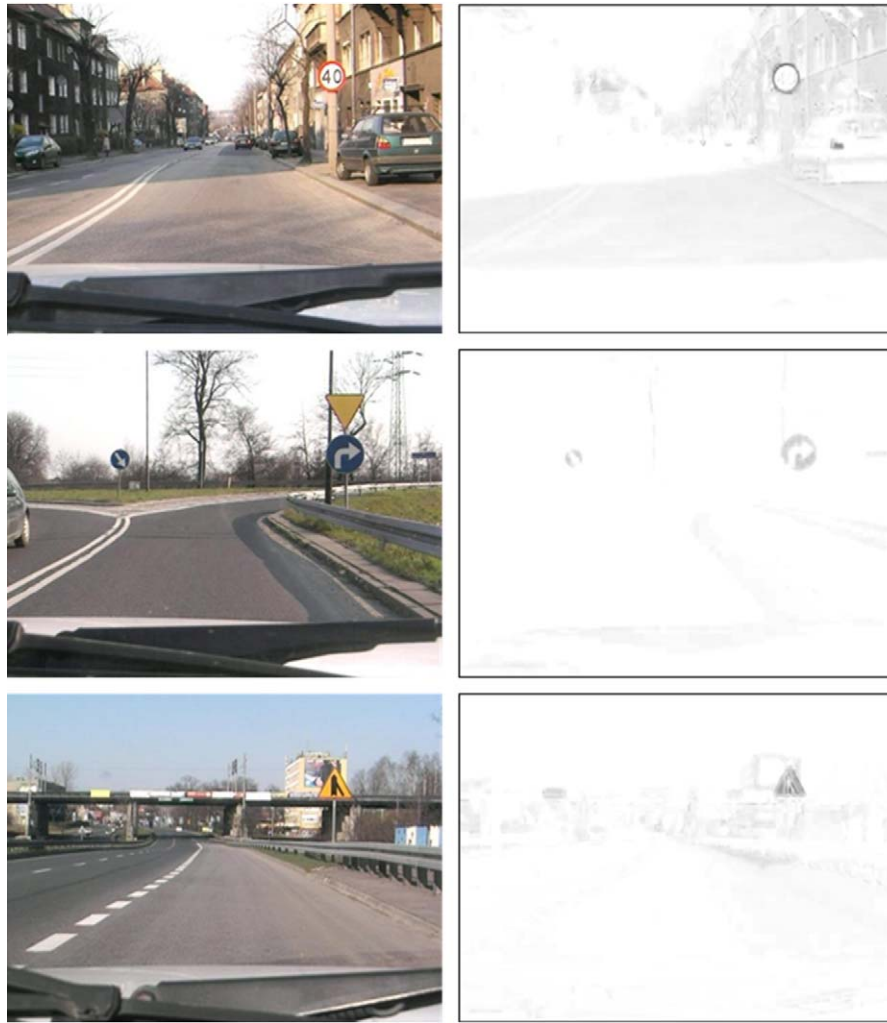


Fig. 2. The effect of filtering of the original RGB images (left) using the transforms defined in (1). Red, blue and yellow colour enhancement, respectively, are shown from top to bottom. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

range of scales. For instance, for a “blue pair” a circular shape detector is triggered to search for the blue instruction signs, e.g. “turn left” or “turn right”, and a square detector is run to detect potential square information signs, e.g. “pedestrian crossing” or “parking place”. It should be noted that the signs considered in the experimental section of this paper are specific to Poland, where the test video sequences were captured. In most countries, for instance in the United Kingdom or in Germany, the cautionary signs are white triangles with distinctive red rim, rather than yellow triangles that are usually reserved for temporary roadwork signs. Adapting the presented detection algorithm to handle such signs is straightforward. The equilateral triangle detector simply has to be run with the gradient and edge images associated with the red colour (and the first transform in Eq. (1)). As in this case the yellow colour is no longer of interest, the preprocessing step is simplified and hence the overall speed of the detector should be increased.

Minimum response thresholds of the aforementioned colour- and shape-specific detectors are tuned using an independent set of reasonably clean static sign images. Optimal setting of each threshold is done by picking the highest value for which an appropriate shape is detected in all examples. As each found candidate has known shape and rim colour, detector serves as a pre-classifier reducing the number of possible templates to analyse at the later stage of the processing to the ones contained in either category. When signs are in

the cluttered background, some number of false positives may be produced. To address this issue, an additional step is taken to verify that the image fragment enclosed in each candidate contour contains pixels of the colour specific to the interior of the signs representing the just determined category.

Once a candidate sign is detected, it is unnecessary to search for it in the consecutive frames in every possible image location. We have employed a Kalman filter [1] to track a sign detected in a previous frame of an input video. An assumption of constant velocity and straight-line motion of a vehicle is made. The state of the tracker is defined by (x, y, s_x, s_y) , where x, y are coordinates of the sign's centre in the image, and s_x, s_y describe its apparent size. Enforcing a constant direction and magnitude of the vehicle's velocity vector imposes certain limitations in the kinematics of the model. In practice, these limitations are not severe because the Kalman filter is only used as a tool for local search region reduction and hence does not directly affect the results of shape detector. Moreover, when the detected candidate sign departs from its hypothetical trajectory due to the inaccurate velocity assumption (e.g. when the vehicle suddenly accelerates, swerves, or brakes), the process error covariance of the filter is increased accordingly. This property of the KF helps maintain good accuracy of the localisation even when the motion of the vehicle is not perfectly smooth in terms of the direction and velocity.

When the traffic sign is for the first time detected, it is sufficiently far from the camera to consider it geometrically undistorted. Although this distortion should theoretically increase when the camera approaches the target, the true scale of the affine signs' deformation in the subsequent video frames is still very small. However, to make our tracker even more accurate, we have augmented it with an additional anisotropic scale corrector. Specifically, when the prior estimate of the state at time t is made by the KF, a particle filter implementing the sampling importance resampling (SIR) algorithm [5] is run to refine the previous scale estimates, s_x , s_y , around the centre of the radial-symmetric shape found by the regular polygon detector.² Each particle is re-weighted based on how much the hypothesised corrected shape's contour encoded by it fits the currently observed gradient orientations and magnitudes. Because only two parameters are optimised, a relatively small number of particles suffice to achieve a satisfactory correction without noticeable computational slowdown. The scale estimates refined by the particle filter, together with the centroid's location captured by the regular polygon detector, are used to update the KF state at time t .

It should be emphasised that the current mean and variance estimates from the Kalman filter are used to locate the centre and the size of the local search region in the next video frame. Within this fragment of the image the same regular shape detection process is repeated as at the time of the initial candidate discovery, but now within a much smaller area and using a single detector instance, appropriate to the already known general sign's category. Therefore, computation has been significantly reduced compared to the exhaustive search over the whole image, which makes it possible for the tracker to run at frame rate.

3. Image representation

Selecting an optimally discriminative feature set for a large number of traffic sign images is a non-trivial task. The simplest choice is probably to extract the local image characteristics at the pre-defined, regularly distributed locations uniform for all signs, as Gao et al. [25] did. Naturally, the drawback of this approach is that the feature locations are chosen manually, irrespective of how much discriminative information the corresponding image fragments carry. Another possibility is to describe each sign by some global numerical characteristics, e.g. moments. However, this technique proves useful when the number of classes to recognise is small and these classes are themselves significantly different from one another. With the increase in the number and similarity of classes, moments and other global shape descriptors become less discriminative, which was confirmed in the preliminary experiments we made using the still camera images of road signs.

We have experimented with several automatic feature selection techniques such as Principal Component Analysis (PCA) and AdaBoost. Details of this experiment are given in Section 6.2. Aiming at retrieving the global variance of a whole dataset, PCA is not capable of capturing features critical to the individual classes. On the other hand, AdaBoost framework is known to provide a way for extracting a compact representation and generating efficient classifiers. However, it is originally designed to solve binary problems and a generalisation to multi-class problems is not straightforward. Besides, a large amount of data is required for AdaBoost training. Collection and preprocessing of such data is a very time-consuming process, with an additional difficulty being caused by the fact that certain road signs occur extremely rarely. Finally, traffic signs are very well defined objects with only small intra-class variability and therefore can be unambiguously represented with clean prototypes. Therefore,



Fig. 3. Sample traffic signs with the high-entropy salient regions marked. Salient regions were extracted using a simplified version of Kadir and Brady's algorithm [16] that allows only square regions. Obtained regions were clustered in the spatial domain. Note that in the case of the two last signs no salient regions were found. This figure is best viewed in colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

a question arises whether there is sufficient justification for learning them from the real-life data.

A separate issue is the choice of the underlying image representation from which salient features could be generated, and the actual meaning of "saliency". We found most of the high-level visual shape descriptors such as moments inadequate to represent the detected candidate road sign images due to the low resolution of these images. On the other hand, low-level, pixel-based methods tend to suffer from high sensitivity to all kinds of geometrical distortions, e.g. shifts and rotations. Histogram-based methods partly overcome this limitation, but in return they lose the valuable information about the spatial arrangement of the individual pixels. In the preliminary experiments our sign detector demonstrated good performance in detecting the relevant sign shapes. However, the match between the hypothetical contour and the true shape in the image is not always perfectly accurate in the noisy, low-resolution video frames, especially those depicting the signs at a considerable distance. These inevitable misalignments caused by the detector prevent the classical pixel-wise and histogram-based representations from being useful and call for more adequate sign representations.

We finally trialled a patch-based approach to describe an observable sign pattern with a collection of salient local features. One of the well-justified meanings of visual saliency, proposed by Kadir and Brady [16], was considered at this point, but the observations below apply to the other methods of this kind, e.g. SIFT [19]. Three kinds of problems were encountered. First, different traffic signs cannot be represented by even roughly equal numbers of salient regions, as some in fact contain nearly nothing visually salient in a sense of large entropy, e.g. "no vehicles" sign or "give way" sign (without the inscription inside) shown in Fig. 3. Second, due to generally small apparent scales of road signs and a very small number of colours characterising them, the meaning of entropy defined over the pixel intensity becomes vague. Different definitions are also problematic unless a large amount of training data is available. Finally, "salient" defined by the entropy within a single sign image does not mean "discriminative" among a group of signs, especially when these signs are similar to one another.

Motivated by [27], we propose in Section 4 an algorithm that extracts for each template sign a limited number of local image regions in which it looks possibly the most different from all other templates in the same category. This way we define an alternative meaning of visual saliency to the one suggested by Kadir and Brady [16]. The extracted discriminative regions are further used for comparing the noisy video frame observations with the idealised road sign templates to make a reliable on-line sign classification. In the rest of this section we first outline the process of converting a raw bitmap image into a more suitable discrete-colour representation and define a *Colour Distance Transform* on which a robust region distance metric is built. Definition of discriminative local regions and the aforementioned dissimilarity metric, as well as a description of a region selection algorithm are postponed to Section 4.

3.1. Colour discretisation and colour distance transform

The tracked road signs are passed on input of the recognition module as rectangular image regions containing the target object

² Regular polygon detector normally captures even slightly distorted road signs.



Fig. 4. Sample images obtained by the sign detector before (above) and after (below) background masking and colour discretisation; 2 bits suffice to encode colours in each image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and, depending on its shape, also background fragments, as depicted in Fig. 4. If there is any in-plane rotation or anisotropic scaling indicated by the detector/tracker (refer to Section 2 for details), it is first compensated. In order to prepare the candidate for classification, the image is scaled to a common size, typically 60×60 pixels for circular and square signs, and 68×60 pixels for triangular signs. Undesirable background regions are then masked out using the information about the object's shape provided by the detector. It is important to note that the full colour spectrum is far more than necessary to identify the pictogram, as the signs contain only up to four distinctive colours per category. Therefore, the candidate images are subject to on-line colour discretisation according to the category-specific colour models learned off-line from a set of training images as follows.

For each category of signs a number of frames are picked randomly from a set of realistic traffic video sequences depicting the respective signs. From the region occupied by a sign in each image we manually pick several pixels representing known named colours and record their RGB values, which are further transformed to CIE XYZ values.³ This is how the training data are constructed. Then, the Expectation Maximisation algorithm [2] is employed to estimate an optimal Gaussian Mixture Model (GMM) [4] for each colour specific to this category. The procedure is restarted several times for the increasing number of randomly initialised Gaussian components and the best model in terms of the mean data likelihood is saved.

In system runtime the appropriate off-line learned GMMs are used to classify each RGB pixel into one of the admissible colours by picking the model that has most likely generated this pixel. On the implementation side, it should be noted that a large speedup of this colour discretisation can be achieved at the cost of higher memory consumption. Namely, the colour models can be used in advance to assign the appropriate colour label to each possible RGB triple, yielding a look-up table with 255^3 entries for each sign category. This way, intensive computation can be avoided by picking the colour labels directly from the memory-stored look-up tables. Sample results of the on-line colour discretisation described above are illustrated in Fig. 4.

Our ultimate goal is to enable robust comparisons between the realistic and the model images in a discrete-colour representation. To this end we know how to rapidly obtain such a representation from the incoming RGB image regions enclosing the detected sign candidates. We also possess the template images where the colour palette is already sparse. To facilitate the aforementioned comparisons, a separate distance transform DT [3] is computed for each discrete colour, giving output similar to this shown in Fig. 5. In DT computation pixels of a given colour are simply treated as feature pixels and all the remaining pixels are treated as non-feature pixels. A (3,4) Chamfer metric [11] is used to approximate the Euclidean

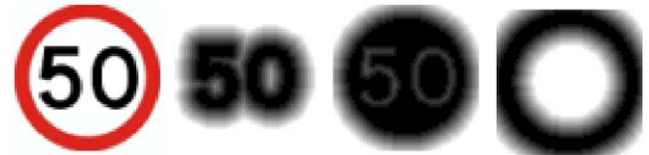


Fig. 5. Normalised Colour Distance Transform (CDT) images. From left to right: original discrete-colour image, black CDT, white CDT, red CDT. Darker regions denote shorter distance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distance between the feature pixels. To emphasise a strong relation to colour, we call this variant of DT a *Colour Distance Transform*.

One practical problem emerges when a given colour is absent in one input image. Namely, within each category of signs there are some that do not contain the colours present in the other. However, for the sake of comparisons, all colour distances must be sensibly represented in each template sign. To reflect the absence of a given colour in CDT, positive infinity is not appropriate as it causes numerical problems. Instead, we introduced a fixed maximum relevant colour distance value $d_{max} = 10$ pixels and normalised the colour-specific DT images by assigning each pixel p in the image I a value defined as:

$$\tilde{d}_{CDT}(I, p) = \begin{cases} \frac{d_{CDT}(I, p)}{d_{max}} & \text{if } d_{CDT}(I, p) \leq d_{max}, \\ 1.0 & \text{if } d_{CDT}(I, p) > d_{max}. \end{cases} \quad (2)$$

3.2. Justification of CDT

The main idea behind the Colour Distance Transform is to define a smooth distance metric to be used in the comparisons between the image of a likely sign observed in the input video and the template images. We want this distance metric, discussed in Section 4.1, to capture the variability of the road signs' appearance. This variability could normally be modelled statistically from a large number of training images. However, collection of a sufficiently large number of such images is difficult as certain signs are in practice extremely rare. In the same time, we would not like to restrict ourselves to a narrow subset of the most popular signs, as many practitioners do.

To justify our distance metric choice, let us consider the potential causes of why the same pictograms may look different from scene to scene. First, there might be physical differences between the same pictograms in different countries. For instance, the digits in the speed limit signs might be painted using different fonts. A good insight into how the traffic signs differ across countries can be found in [30]. From the point of view of possible applications, it would probably be the most practical to consider a recognition system trained on the country-specific sign database and able to re-train itself instantly for operation in other countries, or simply switch a pre-loaded country-specific classifier, when needed. However, even within the same country, minor physical differences

³ CIE XYZ colour appearance model is used in preference to the raw RGB space because it is device-independent.

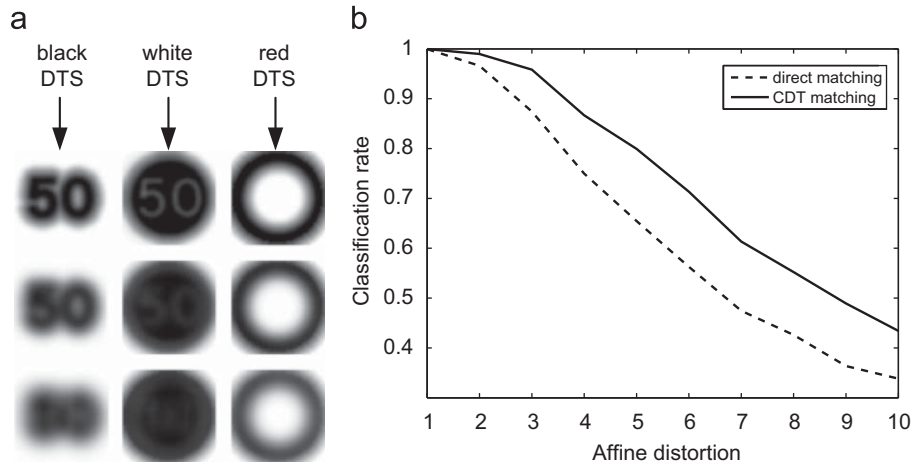


Fig. 6. Experimental evaluation of Colour Distance Transform. (a) Colour distance maps: black, white, and red, respectively, obtained for a sample speed limit sign in the experiment involving generation of 1000 random affine deformations. Each parameter of the transformation: translation, x , y , rotation around the centroid, θ , scale, s_x , s_y , and shear, h_x , h_y , was perturbed according to a the clipped normal distribution. Standard deviations of these distributions were: $\sigma_x = \sigma_y = 1$ px, $\sigma_\theta = 1^\circ$, $\sigma_{s_x} = \sigma_{s_y} = 0.02$, $\sigma_{h_x} = \sigma_{h_y} = 0.02$ (top row), $\sigma_x = \sigma_y = 3$ px, $\sigma_\theta = 3^\circ$, $\sigma_{s_x} = \sigma_{s_y} = 0.02$, $\sigma_{h_x} = \sigma_{h_y} = 0.02$ (central row), $\sigma_x = \sigma_y = 5$ px, $\sigma_\theta = 5^\circ$, $\sigma_{s_x} = \sigma_{s_y} = 0.05$, $\sigma_{h_x} = \sigma_{h_y} = 0.05$ (bottom row). (b) Correct classification rate for a 42-class Polish cautionary signs problem as a function of image distortion, obtained using two discrete image comparison methods. A simple nearest neighbour template matching classifier was used. In the first comparison method (dashed line), each distorted image was compared to the undistorted template signs by counting the numbers of spatially corresponding pixels having unmatching colours. In the second method (solid line), distance between each compared pair of images was measured pixel-wise using the CDT-based metric. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

between the same road sign classes exist. Fortunately, in the practical realisations of the traffic sign recognition systems the signs are usually captured and analysed when still being at a significant distance from the camera. In the resulting low resolution imagery the high-frequency details and the vendor-related pictogram differences become insignificant.

Other sources of intra-class variability are related to the target perception process and the nature of its imperfections. For example, the appearance of the same type of sign varies due to changing illumination or different colour/reflectance properties of the surface of signs' boards. This is to certain extent accounted for by the GMM-based colour binning, which was explained in Section 3.1. This method can model multimodal colour distributions and is hence robust to substantial illumination changes, as long as the colour information is not completely destroyed by the factors such as strong incident light, reflections, or deep shade. Finally, the same signs may look slightly different due to small viewpoint changes. We show that this problem is addressed by the *Colour Distance Transform* as it enables modelling the distribution of a discrete-colour appearance of traffic signs under small affine transformations without recourse to the massive volumes of natural data.

To make the last point more convincing, we have done two simple experiments. First, from each discrete-colour image of a chosen template sign $n = 1000$ random affine transformations were generated. All parameters of the transformation matrices: translation, x , y , rotation around the centroid, θ , scale, s_x , s_y , and shear, h_x , h_y , were drawn from the clipped normal distributions with appropriately low standard deviations to ensure the generated distortions were realistically small. In this experiment an alternative method of constructing the colour distance maps was evaluated. Namely, for each pixel we counted how many times this pixel was not of a given colour in each distorted image and divided it by n . Fig. 6a illustrates the resulting frequencies obtained for different values of standard deviation of the affine transform parameters. Clearly, these images very much resemble our CDT images (see Fig. 5 for comparison) when the distortion parameters are appropriately chosen.

In the second experiment we directly compared accuracy of template matching under small transformations using: (1) a distance metric based on the co-occurrence of discrete colours in both im-

ages, and (2) a distance metric based on CDT. To clarify, in the first method simply a fraction of the spatially corresponding pixels having different colours in both discretised images was calculated. In this experiment each of 42 model cautionary signs was artificially distorted 100 times. Each such distorted image was compared to each undistorted model, i.e. to one template representing the same sign and 41 templates representing the remaining signs. The standard deviations of the affine transformation parameter distributions, the same as those used in the first experiment, were gradually increased, starting from: $\sigma_x = \sigma_y = 0.5$ px, $\sigma_\theta = 0.5^\circ$, $\sigma_{s_x} = \sigma_{s_y} = 0.005$, $\sigma_{h_x} = \sigma_{h_y} = 0.005$, and with step: $\Delta\sigma_x = \Delta\sigma_y = 0.5$ px, $\Delta\sigma_\theta = 0.5^\circ$, $\Delta\sigma_{s_x} = \Delta\sigma_{s_y} = 0.005$, $\Delta\sigma_{h_x} = \Delta\sigma_{h_y} = 0.005$. Correct classification rates obtained are shown in Fig. 6b.

Results of the above experiments suggest that the Colour Distance Transform is suitable for comparing pairs of discretised images of traffic signs affected by minor affine transformations. Smooth distance metric defined over CDT clearly outperforms simple pixel-wise discrete-colour matching. Together with the proposed colour discretisation technique, CDT becomes a good alternative to the pose- and illumination-invariant traffic sign appearance modelling. It is superior to the data-driven methods in that, with the exception of the images needed for GMM colour classifier training, it does not require the realistic traffic sign images.

4. Feature selection

As a CDT-based smooth distance metric is available, discrete-colour images of traffic signs can be compared pixel by pixel, without risking serious recognition rate degradation caused by small image misalignments. However, our intuition is to reduce the dimensionality of the feature space not only by suppressing redundant colour information, but also by selecting only those fragments of each pictogram that are really unique for the sign it is representing.

4.1. Discriminative local regions

A complete space of local regions is obtained by subdividing the image into small, regularly spaced, non-overlapping square blocks

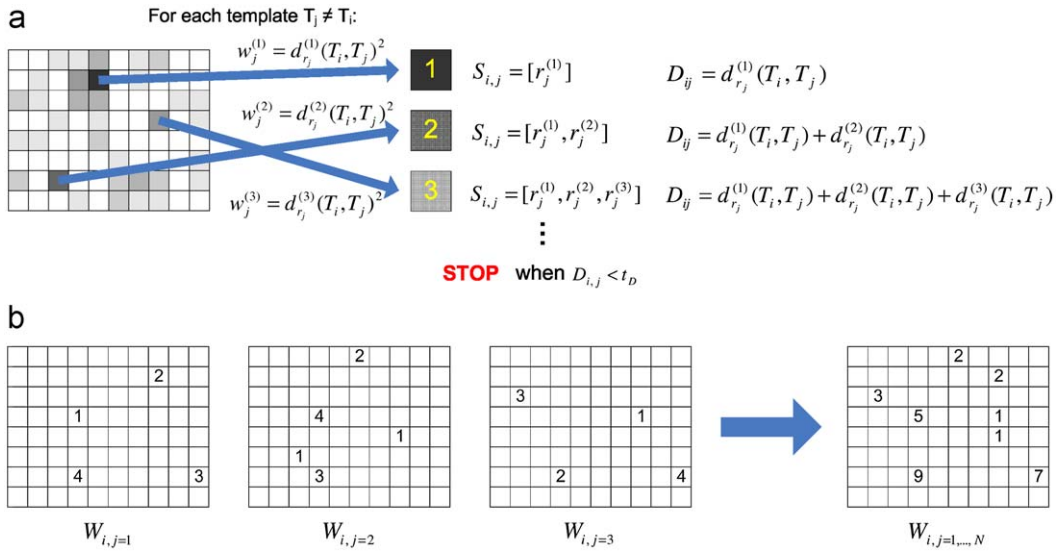


Fig. 7. Region selection algorithm: (a) construction of the local region partial ranking characterising differences between templates T_i and T_j , (b) merging three sample partial rankings; the partial weights of each relevant region are given as whole numbers only for illustration purposes.

of size $m \times m$ pixels. Typically, we use $m = 4$ pixels which yields the total of 225 regions for non-triangular sign images of size 60×60 pixels ($60/4 \times 60/4 = 225$) and 255 regions for triangular sign images of size 68×60 ($68/4 \times 60/4 = 255$). Within each region r_k dissimilarity between the images I and J can be calculated using the discrete-colour image of I and the normalised CDT images of J by averaging the pixel-wise distances:

$$d_{r_k}(I, J) = \frac{1}{m^2} \sum_{t=1}^{m^2} \psi_{CDT}(I, J, p_t), \quad (3)$$

where for each pixel p_t contained in the region, distance $\psi_{CDT}(I, J, p_t)$ is picked from the appropriate normalised CDT image of J , depending on the colour of this pixel in I . Let us also denote by $\hat{d}_S(I, J)$ and $\hat{d}_{S,W}(I, J)$ a normal and weighted average local dissimilarities between the images I and J , computed over regions $r_k \in S$ (weighted by $w_k \in W$):

$$\hat{d}_S(I, J) = \frac{1}{|S|} \sum_{k=1}^{|S|} d_{r_k}(I, J), \quad (4)$$

$$\hat{d}_{S,W}(I, J) = \frac{\sum_{k=1}^{|S|} w_k d_{r_k}(I, J)}{\sum_{k=1}^{|S|} w_k}. \quad (5)$$

As CDT images for the model signs are pre-computed, any on-line local-region comparison between the observed and the template images can be made extremely fast.

4.2. Region selection algorithm

Assuming pre-determined category of signs $C = \{T_i : i = 1, \dots, N\}$ and a candidate image \mathbf{x}_j , our goal is to determine the class⁴ of \mathbf{x}_j

by maximising posterior:

$$p(T_i | \mathbf{x}_j, \theta_i) = \frac{p(\mathbf{x}_j | T_i, \theta_i) p(T_i)}{\sum_{k=1}^N p(\mathbf{x}_j | T_k, \theta_k) p(T_k)}. \quad (6)$$

We think that uniform feature sets are inadequate for traffic sign recognition. Some signs can be told apart by just a single distinctive pictogram element, while other need to be analysed in more detail to be distinguished from other similar signs. Our objection to using a uniform feature space for classification makes us envisage different model parameters $\theta_i = (\mathbf{I}_i, \mathbf{W}_i)$ for each template T_i . \mathbf{I}_i denotes an indexing variable determining the set S_i of regions to be used and \mathbf{W}_i is a vector of relevance corresponding to the regions $r_k \in S_i$ selected by \mathbf{I}_i . In order to learn the best model parameters θ_i^* , the following objective function is maximised:

$$O(\theta_i) = \sum_{j \neq i} \hat{d}_{S_i}(T_j, T_i). \quad (7)$$

In other words, the regions best characterising a given sign are obtained through maximisation of the sum of local dissimilarities between this sign's template and all the remaining signs' templates.

In presence of model images only, each average set dissimilarity term $\hat{d}_{S_i}(T_j, T_i)$ as a function of the number of discriminative regions in S_i is necessarily monotonically decreasing. This is because the subsequent regions added are increasingly less dissimilar and hence give smaller contribution to this average. As a result, typically there would be just a few good regions or even a single best region maximising equation (7). In practice, such sign descriptors are unlikely to work well for the noisy video where more support in terms of the number of image patches to match in each frame is required to make a reliable discrimination. Therefore, to balance the discriminative power and the reliability, our objective function is iteratively degraded up to the specified breakpoint, yielding a representation which is more dense and thus more useful in a real data context.

Similarly to Paclík et al. [27], in the model training stage we have adopted elements of a sequential forward search strategy, a greedy technique from the family of floating search methods [6]. However, both approaches differ significantly in two main aspects. First, we think that learning the signs from the real-life images might not be worth the effort required as the publicly available templates seem to sufficiently characterise the appearance of the respective

⁴ Throughout the following sections of this paper by using the term “class” we always mean an unambiguous semantic identity of a sign which always maps to a single template. No generic, higher-level classes are considered, e.g. general “speed limit”. The only exception are the four main categories: instruction signs, prohibitive signs, cautionary signs, and informative signs, but these are always referred to as “categories”.

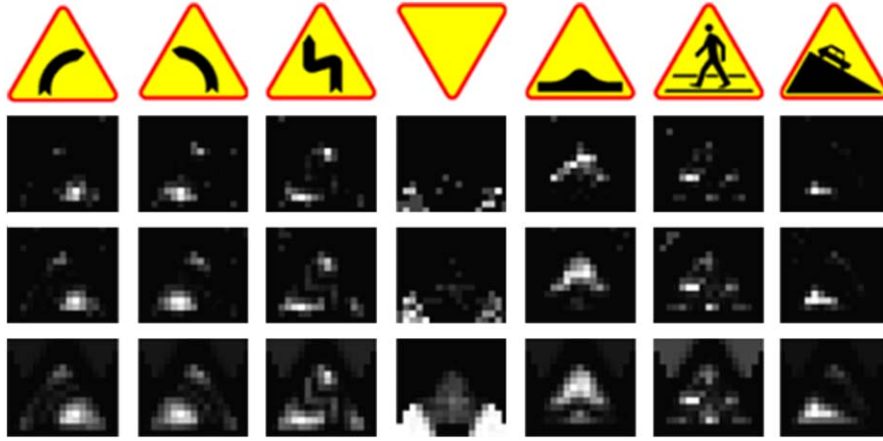


Fig. 8. Sample triangular template images (1st row), and 4×4 -pixels discriminative regions obtained for the parameter $t_D = 2.0$ (2nd row), $t_D = 5.0$ (3rd row), and $t_D = 50.0$ (4th row). Brighter regions correspond to higher dissimilarity.

classes. Second, we believe that the possible within-class appearance variability may well be accounted for by a robust distance metric, as the one introduced in Eqs. (3)–(5), instead of being learned. Our implementation is outlined in Algorithm 1.

Algorithm 1. Discriminative region selection.

Input: sign category $C = \{T_j : j = 1, \dots, N\}$, target template index i , region pool $R = \{r_k : k = 1, \dots, M\}$, dissimilarity threshold t_D

Output: ordered set S_i of regions to discriminate between template T_i and all other templates, ordered set W_i of weights corresponding to the regions from S_i

- 1: initialise an array of region weights $W = \{w_k : w_k = 0, k = 1, \dots, M\}$
- 2: **for each** template $T_j \in C, j \neq i$ **do**
- 3: sort R by decreasing dissimilarity $d_{r_j}(T_i, T_j)$
- 4: initialise ordered region set S_{ij} , and the corresponding weight set W_{ij} , characterising the dissimilarity between templates T_i and T_j , with the first region from R and its weight, respectively:
 $S_{ij} = [(r_j^{(1)})]$, $W_{ij} = [w_j^{(1)}]$, where $w_j^{(1)} = d_{r_j^{(1)}}(T_i, T_j)^2$
- 5: initialise region counter $l = 1$
- 6: initialise the total dissimilarity, D_{ij} , between templates T_i and T_j : $D_{ij} = d_{r_j^{(1)}}(T_i, T_j)$
- 7: **while** $D_{ij} < t_D$ and $l < M$ **do**
- 8: increment region counter $l = l + 1$
- 9: set weight of the new region to: $w_j^l = d_{r_j^{(l)}}(T_i, T_j)^2$
- 10: add region $r_j^{(l)}$ to S_{ij} and weight $w_j^{(l)}$ to W_{ij}
- 11: update D_{ij} : $D_{ij} = D_{ij} + d_{r_j^{(l)}}(T_i, T_j)$
- 12: **end while**
- 13: **for each** region r_k such that $r_k \in S_{ij}$ **do**
- 14: update region weight: $w_k = w_k + w_j^{(l)}$,
where $w_j^{(l)}$ is the weight of region r_k in W_{ij}
- 15: **end for**
- 16: **end for**
- 17: build the target region set S_i and the target weight set W_i : $S_i = \{r_k : w_k > 0\}$, $W_i = \{w_k : w_k > 0\}$

A given template sign is compared to each of the remaining templates. In each such comparison the algorithm loops until the appropriate number of local regions is selected. At a given step of the loop the most dissimilar region is fixed and removed from the pool of available regions. Moreover, at the k -th step the distance between the considered image and the image being compared to is

measured with respect to the joint set comprised of the new k -th region and all previously selected regions. At the end of the loop a list of regions is produced, together with the list of corresponding weights reflecting the discriminative power of these regions. Each pairwise region set build-up is controlled by a global threshold, t_D , specifying the maximum allowed cumulative dissimilarity between any pair of templates being compared. Such a definition of STOP criterion ensures that the same amount of dissimilarity between any pair of templates is incorporated in the model. This in turn allows us to treat different sign classes as directly comparable, irrespective of the actual number of local regions used to characterise them. A single comparison between the interest template T_i and template T_j , $j \neq i$, is schematically depicted in Fig. 7a. The final region set for each class is constructed by merging the pair-specific subsets, as shown in Fig. 7b. It is reflected in the region weights carrying the information on how often and with what importance each particular region was selected.

For each sign the above procedure yields a set of its most unique regions. It should be noted that in the final step, depending on the actual dissimilarity threshold specified, certain number of regions will be found completely unused, and hence discarded. An example output of the proposed region selection algorithm is depicted in Fig. 8. Obtained discriminative region maps clearly show that different signs are best distinguishable in different fragments of the contained pictogram. It can also be seen that although the same value of global parameter t_D was used, different numbers of relevant regions were found.

In absence of realistic images of traffic signs, it is generally hard to choose the optimal value of t_D . We have conducted an experiment that only gives a clue on how the value of this threshold influences discriminative capability of the road sign classifier. In this experiment a clean template image representing a given cautionary sign was geometrically distorted 100 times in a random way, as previously introduced in the experimental evaluation of CDT (Section 3.2). Each distorted image was then matched against all undistorted templates and the percentages of correct matches were recorded. They were further averaged over all 42 types of cautionary signs tested. Deviations of the Gaussian-distributed affine transform parameters were chosen as follows: $\sigma_x = \sigma_y = 3$ px, $\sigma_\theta = 3^\circ$, $\sigma_{s_x} = \sigma_{s_y} = 0.02$, $\sigma_{h_x} = \sigma_{h_y} = 0.02$, which simulates similar affine deformations than those the real signs captured by our regular polygon detectors are typically subject to. For the above fixed geometrical transformation parameters, we varied the value of t_D to observe how it influenced the correct classification rate of the template matching classifier. The result is presented in Fig. 9.

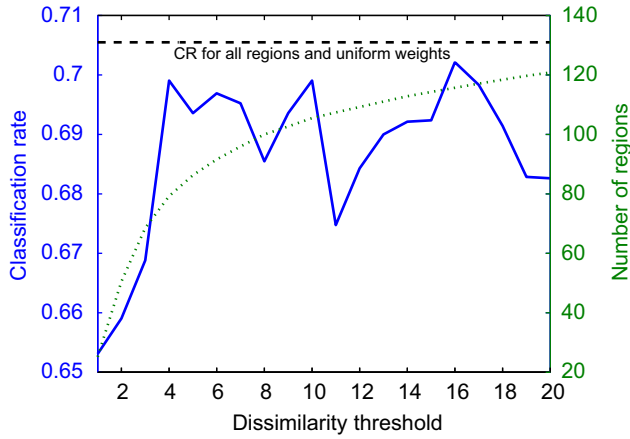


Fig. 9. Recognition rate of a nearest neighbour classifier employed to classify geometrically distorted template road sign images. Parameters of the normal distribution controlling affine perturbation of the images were fixed and the parameter t_D was a variable. Apart from the correct classification rate also the average number of regions in the target discriminative representation of sign is shown as a function of t_D . Recognition rate obtained when using all available unweighted regions is marked with dashed line.

The recognition rates of the aforementioned classifier seem to indicate that it is best not to reduce the number of discriminative local regions at all, i.e. simply compare the entire images using CDT-based metric. However, as for only one-third of all available regions the recognition performance is only minimally worse than when using all regions, dimensionality reduction is still worthwhile. Despite the above observation, it is risky to assume that more dense representations always imply better discriminative power of the classifier. In the above experiment only the possible geometrical transformations were simulated, but the influence of other factors (e.g. changing illumination, motion blur, contamination of the image with the incidental background fragments) on the classifier's performance as a function of the number of discriminative regions used may be different. Tuning of the dissimilarity threshold based on a small number of real traffic sequences captured from a moving vehicle, which we adopt, is more adequate in this case. More details on this parameter tuning are given in Section 6.1.

5. Temporal classifier design

The proposed road sign classifier distinguishes between the sign classes contained in a category pre-determined in the detection stage, based on the discriminative feature representation unique for each particular sign. For simplicity, two assumptions are made: (1) the dissimilarity between each sign and all other same-category signs is Gaussian-distributed in each local region and independent of the dissimilarities in all other regions characterising this sign, and (2) class priors $p(T_i)$ are equal. In such a case Maximum Likelihood theory allows us to relate the maximisation of likelihood $p(\mathbf{x}_j|T_i, \theta_i)$ to the minimisation of weighted distance $\hat{d}_{\mathbf{S}_i, \mathbf{W}_i}(\mathbf{x}_j, T_i)$. Therefore, for a known category $C = \{T_i : i = 1, \dots, N\}$ and observed candidate \mathbf{x}_t at time t , the winning class $L(\mathbf{x}_t)$ is determined from (8):

$$L(\mathbf{x}_t) = \arg \max_i p(\mathbf{x}_t|T_i, \theta_i) = \arg \min_i \hat{d}_{\mathbf{S}_i, \mathbf{W}_i}(\mathbf{x}_t, T_i), \quad (8)$$

where the elements of the region set \mathbf{S}_i and the corresponding weights in \mathbf{W}_i denote the ones learned in the training stage for the template T_i .

When a series of observations from a video sequence is available, it is reasonable to integrate the classification results through the whole sequence over time, instead of performing individual classifications. Hence, at a given time point t our temporal integration

scheme attempts to incorporate all the observations made since the sign was for the first time detected until t . Denoting observation relevance at time t by $q(t)$ and assuming independence of the observations from consecutive frames, the classifier's decision is determined by:

$$L(\mathbf{X}_t) = \arg \min_i \sum_{k=1}^t q(k) \hat{d}_{\mathbf{S}_i, \mathbf{W}_i}(\mathbf{x}_k, T_i). \quad (9)$$

Usually, the number of frames where the sign is being tracked and recognised is in between 20 and 60, depending on the size of the sign, its location in the scene, and the velocity of the vehicle. We have observed that the signs detected in the early frames are often inaccurately delimited and contain blurred pictograms due to the low image resolution. Also, as colours tend to look paler when seen from a considerable distance, previously discussed colour discretisation exposes severe limitations, unless performed when the candidate sign has already grown in size in the image plane. To address this problem, we adopt the exponential observation weighting scheme from [24] in which relevance $q(t)$ of the observation at time t depends on the candidate's age (and thus size):

$$q(t) = b^{t_{last} - t}, \quad (10)$$

where $b \in (0, 1]$ and t_{last} is the time point when the sign is for the last time seen. The optimal value of parameter b is typically in between 0.7–0.9 which effectively makes the ultimate decision of the classifier mostly dependent upon the last 5–10 observations. This strategy may appear to be losing a large amount of information gathered early in the observation process, but has been experimentally shown to provide the best recognition accuracy.

6. Experiments

To evaluate our traffic sign recognition system, experiments were performed on the real data collected on Polish and Japanese roads. In Section 6.1 we first test our traffic sign recognition system on realistic video captured from a moving vehicle. In Section 6.2 a comparative evaluation of the classifier based on the proposed discriminative local region representation of traffic signs is presented. This classifier is tested on static, low-resolution sign images and compared to the alternative techniques based on PCA and a modified AdaBoost algorithm.


6.1. Overall system performance evaluation

To test the proposed traffic sign recognition system as a whole, a number of real traffic video sequences were captured from a moving vehicle on Polish roads at different times of the year: February, April, June, November, and December. A JVC GR-X5EK DV camcorder mounted in front of the windscreen was used for this task. Its lens was adjusted at the lowest available focal length of $f = 3.2$ mm. The vehicle's velocity varied depending on the traffic situation. It was usually in between 40 and 70 km/h, and never exceed 100 km/h. Test video resolution was 640×480 pixels and its content depicted the total of 210 signs in urban, countryside, and motorway scenes. All sequences we captured in natural daytime lightning, with some signs appearing in shade and in the cluttered background. Due to extreme rarity of certain road signs, the ones in the test data represent only a part of the whole gamut recognised by our system. The detailed breakdown has been provided in Table 1.

For optimal setting of the unknown model parameters, we considered two auxiliary datasets. The first set consisted of 200 images of road signs (50 per category), cropped from various traffic scenes. We used this dataset to (1) adjust the minimum response thresholds

Table 1

Breakdown of the road sign class occurrence in the test data.

	33		6		4		4		3		2		1		1
	22		6		4		4		3		1		1		1
	16		5		4		4		3		1		1		1
	12		5		4		4		3		1		1		1
	7		5		4		3		3		1		1		1
	6		5		4		3		2		1		1		1

The signs not listed were not present in the dataset.

Table 2Detection rates and recognition rates obtained for different values of dissimilarity threshold t_D .

	t_D	RC (55) (%)	BC (25) (%)	YT (42) (%)	BS (13) (%)	Overall (135) (%)
Detected	–	85.3	100.0	95.6	89.6	92.9
	1.0	82.8	92.1	84.9	90.7	87.8
	2.0	93.1	94.7	88.7	90.7	90.3
Recognised	5.0	93.1	97.3	78.8	81.4	85.6
	20.0	89.7	97.3	77.6	79.1	84.6
	All regions	82.6	91.2	66.1	74.3	76.2
	Best	93.1	97.3	88.7	90.7	91.2
Detected and recognised	Best	79.4	97.3	84.8	81.3	85.3

All recognition rates were determined from only these signs that were correctly detected. The third row from the bottom shows the recognition rates obtained when all possible, uniformly weighted local regions were used for image comparison, as though no feature selection was performed. The next to last row contains the recognition rates recorded for the best (trained) setting of t_D -s. In the last row the overall detection and recognition rates are given (for the best classifier setting), i.e. those obtained by multiplying the percentages of the detected and correctly classified signs. The numbers of classes in each category: red circles (RC), blue circles (BC), yellow triangles (YT), and blue squares (BS) are given in parentheses in the column headers.

of the regular shape detectors (see Section 2), and (2) adjust the dissimilarity thresholds t_D , independently for each category of signs.⁵ The second dataset consisted of additional 20 video sequences used for determining the best setting of the temporal weight base b in (10). Optimal value of this weight was found after fixing all category specific t_D -s. It was done by maximising the mean ratio of the cumulative distance between the tracked sign candidate and the best-matching (and correct) template to the cumulative distance between this candidate and the second best-matching template, calculated in the last frame where the tracked sign was seen in the scene.

Table 2 illustrates detection and classification rates obtained for all available test sequences with the model parameters tuned as above described.⁶ To illustrate the influence of the dissimilarity thresholds on the classification rates, the experiments were repeated for varying values of these thresholds. Results were compared with the performance of the classifier generated based on: (1) the exhaustive image comparison (as though no region selection was performed), and (2) image comparison with respect to the sign representations obtained for the optimal setting of t_D -s, determined from an independent dataset, as mentioned above. To visualise the error distribution across the classes, the individual-sign classification results are shown in Table 3.

⁵ Although using this dataset contradicts the idea of learning from model images only, which is one of the major claims of this paper, the road sign images in this dataset do not need to strictly represent all classes. This way the fundamental problem of acquiring the images of very rare signs is avoided.

⁶ It is assumed throughout this paper that the correct classification takes place when the correct template is assigned the smallest cumulative distance in the video frame where the candidate sign is entirely seen for the last time.

As seen in Table 2, obtained classification error rate does not exceed 9%, making our method comparable to the best state-of-the-art approaches. However, it should be noted that our template database contains significantly more signs than in any of the previous studies. Therefore, direct comparison with the alternative methods is not possible. Repetitions of the experiment for different values of dissimilarity threshold revealed that for each category of signs the optimal classifier's performance is achieved for a close to minimum value of this threshold. This stays in contrast with the results of the experiment described in Section 4.2 and suggests that artificial distortion of the idealised sign images is insufficient to tune the t_D parameters. The following observation is vital at this point. The optimal threshold for each category must strike a balance between the two: maximising template signs' separability and the reliability of the obtained dissimilarity information in the real-data context. Very low threshold values lead to the separation of very few good regions for a particular model sign. However, such sparse information may not be sufficiently stable to correctly classify a possibly distorted, blurred, or occluded object in a video frame. Very high 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 signs simply look more similar to one another and are hence more difficult to tell apart, at an additional cost of the more intense computation.

The regular shape detector was set to capture circles/regular polygons of radius in between 15 and 25 pixels inclusive.⁷ Most failures were caused by the insufficient contrast between a sign's boundary

⁷ By radius we mean the radius of the minimum bounding circle enclosing the shape.

Table 3

Individual classification results per sign class.

Green crosses mean correctly classified instances, red crosses mean misclassified instances. Only the correctly detected instances are shown in the table. This table is best viewed in colour.



Fig. 10. Sample pairs of very similar signs that are sometimes confused by the classifier.

and the background, especially for pale-coloured and shady signs. In a few cases this low contrast was caused by the poor quality of the physical target objects rather than their temporarily confusing appearance. Single detection errors emerged when two signs were mounted closely on one pole. In this particular situation candidate objects may be confused with each other, as the local search region of one candidate always contains at least part of its neighbour. Besides, in many cases the shape detector does not yield a perfect contour fit in every video frame. This inaccuracy, internally caused by the presence of confusing edge pixels around the true sign's boundary, becomes significant in the cluttered scenes or when the sign is visible in the low-contrasting background. Difficulty of the scene also affects the system's running speed. With all the category-specific detectors enabled, our implementation can typically run at 25–30 fps. In “favourable” scenes with a dull, greyish background and a single sign present, the system can work at frame rate. Such processing speed is also achieved when the video resolution is decreased to 400×300 pixels. However, in certain situations performance of the system can be significantly degraded, particularly when multiple signs are being tracked at once in high visual clutter.

After closer investigation we observed that approximately one in three classification errors resulted from the confusion between the nearly identical classes, e.g. these shown in Fig. 10. Differences between such signs were found difficult to capture, resulting sometimes in the correct template receiving the second best score. Certain number of misclassifications were caused by the motion blur or inaccurate sign detection, which were in turn a result of car vibration affecting stability of the camera mount. It is common when a vehicle moves on an uneven road surface. Colour binning appeared to be relatively resilient to variations of illumination, leading directly to failure in only several cases when the signs were located in a very

shady area or were themselves of poor quality. This can be a proof of usefulness of Gaussian Mixture colour modelling. In a few cases the vehicle was moving directly towards the bright sun which made it difficult for the driver to recognise the sign's pictogram with his eyes due to destroyed colour appearance information. The corresponding sequences were treated as too challenging and were therefore not used for testing. Remaining failures can be attributed to the imperfections of the detector. For instance, some signs' pictograms consist of edges that may actually be easier to detect than the boundary. This may cause the detected shape to appear clipped.

As indicated in Section 5, fitting regular contour to the observed sign pattern and subsequent colour discretisation are usually less accurate in the early video frames. Extensive experiments have shown that frequently the correct decision is developed by the classifier from just a few last frames where the sign's shape and colours are the most unambiguously determined. This fact provides a good justification for our exponential observation weighting used to promote the most recent measurements. Apparently, the classification accuracy with weighting enabled is by 10–20% higher, depending on the weight base b used. Fig. 11 gives an example of how the temporal weights influence the final score. In the sample charts a ratio of the cumulative distance from the best-matching template to the cumulative distance from the second best-matching template is shown. It is clear that the weighting shifts this ratio towards the desired lower bound of the interval (0, 1).

One limitation of the presented recognition system is that it does not feature any ambiguity rejection mechanism. Apparently, when certain signs appear in the input video, the confidence of the decision is very low due to the existence in the template database a template which is very similar to the correct template. Pairs of such very similar templates have been shown in Fig. 10. A straightforward way of eliminating this ambiguity is thresholding the aforementioned ratio of the cumulative distance from the best-scored template to the cumulative distance from the second best-scored template. However, in presence of signs from Fig. 10 (and other signs of this kind) in the scene, this ratio will always be very close to unity, regardless of how clear the sign is and how accurately it is tracked. In the same time, in presence of other, globally more unique signs, it will smoothly decrease towards zero, as desired. This implies that using a single,

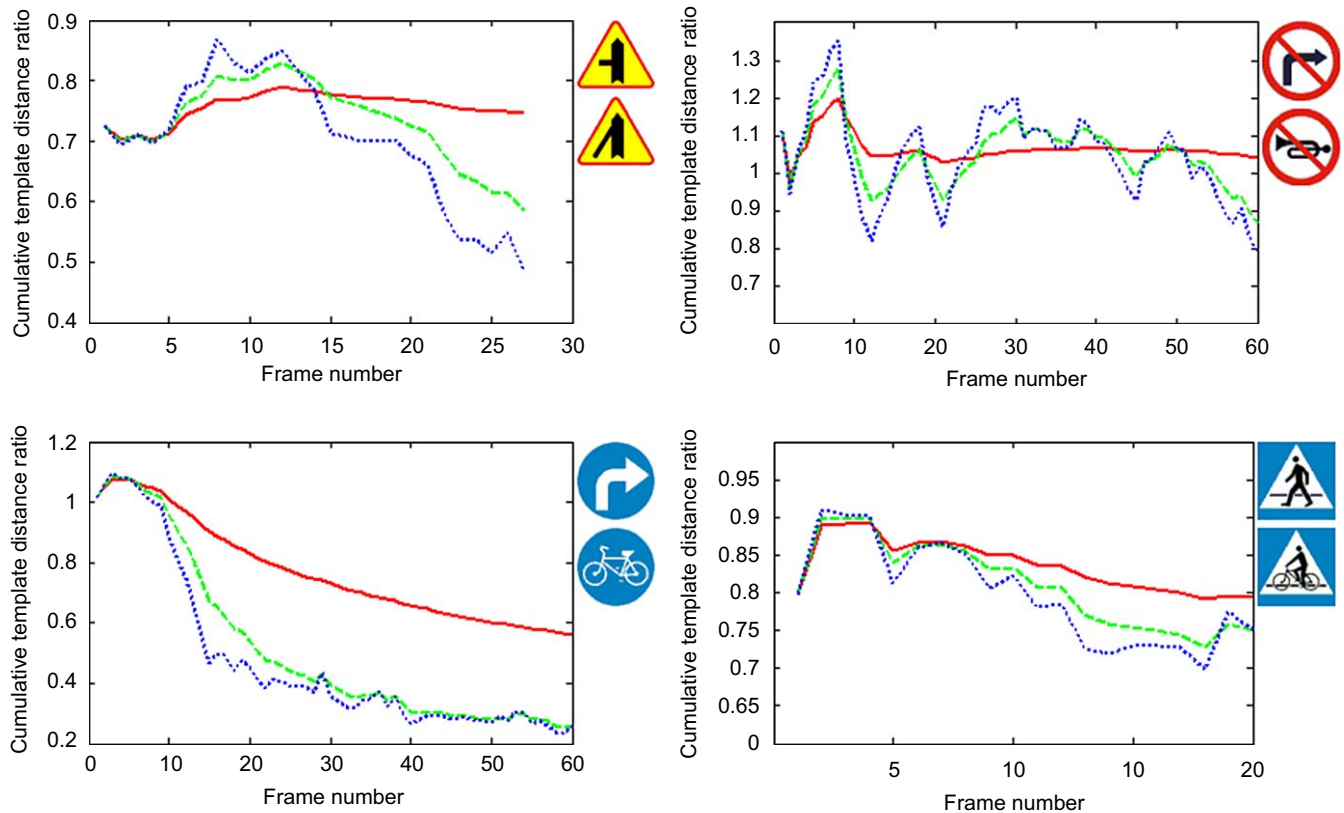


Fig. 11. Classification of signs over time. Ratio of the cumulative distance from the best-matching template (upper sign next to each chart) to the cumulative distance from the second best-matching template (lower sign next to each chart) is marked with a solid red line. The same but temporally weighted cumulative distance ratio is marked with a green dashed line for the weight base $b = 0.8$, and a blue dotted line for the weight base $b = 0.6$. This figure is best-viewed in colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

globally optimal value of “ambiguity threshold” is not supposed to work in all possible situations. If any of the abovementioned signs appeared in the input video, cumulative distance ratio thresholding would classify most detected instances as “not sure”.

Solving the above problem in a more robust way is beyond the scope of the current implementation of our system. One possible direction could be to learn the class-specific sign representations based on the *one-vs-one* dissimilarity maximisation, together with the currently used representations learned via maximising the *one-vs-all* dissimilarity. Such extra representations could be used dynamically (when needed) for comparisons between the actual observation and each of the two most confusing templates. This could resolve at least those confusions involving two very similar pictograms.

6.2. Comparison to PCA and AdaBoost

For comparative analysis of our feature selection algorithm we have performed a separate experiment in which different methods of inferring a discriminative representation of road signs were used. The aim of this experiment was to support our intuition that with the proposed image representation and feature selection methods, a robust, discriminative classifier can be learned easily from the clean template images. In the same time the resulting classifier is no worse than those generated using well-known data-driven techniques such as PCA or AdaBoost. On input we considered a dataset of 13 287 static images (4251 training and 9036 test) extracted from several pre-labelled traffic video sequences. These images represented 17 classes of Japanese traffic signs and were of lower quality than those used in the dynamic recognition experiment. These classes were not equally

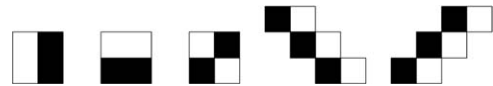


Fig. 12. Haar wavelet features used within the AdaBoost framework in the comparative evaluation of our discriminative feature selection method.

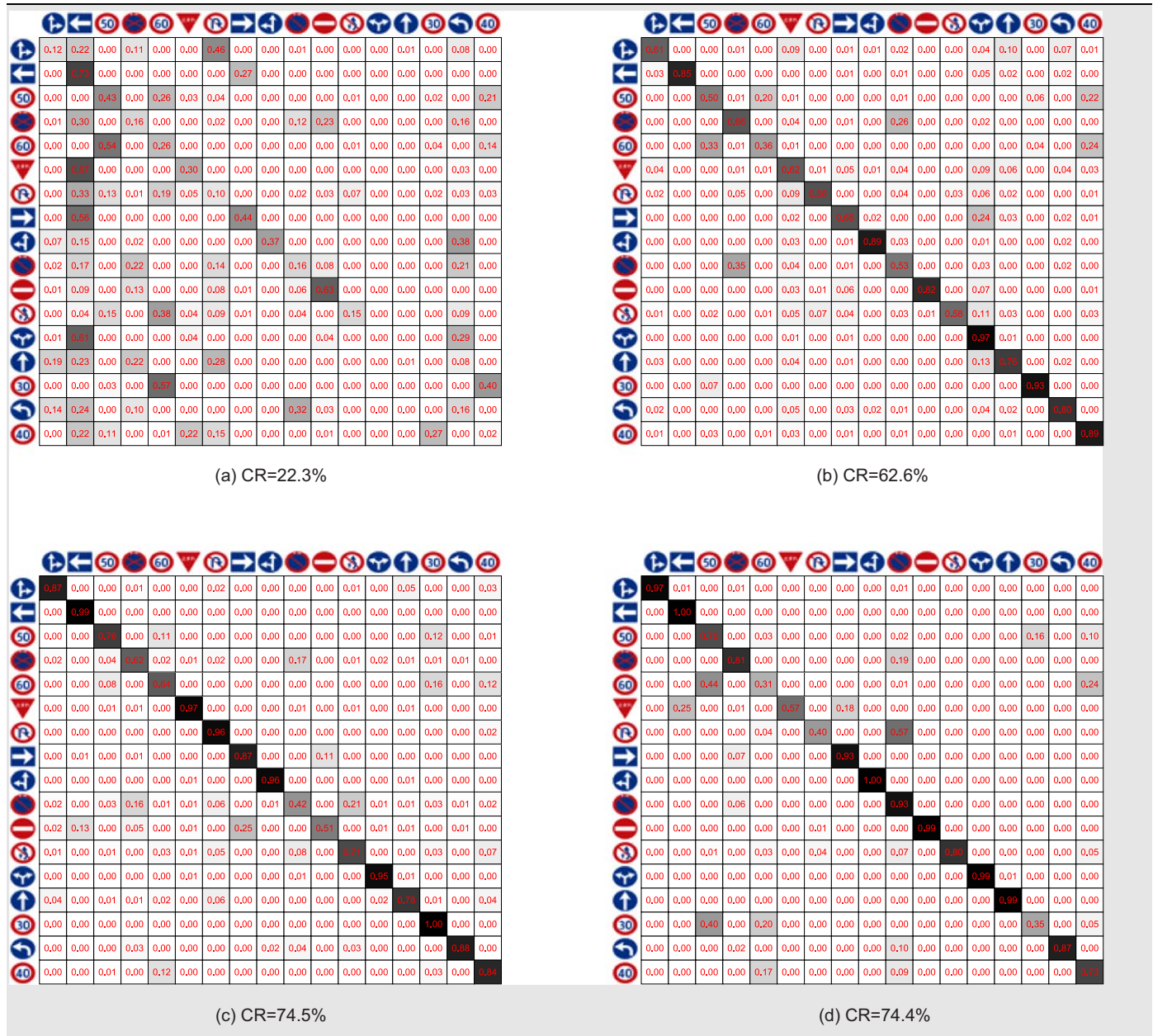
represented in the data, which reflects the fact that certain signs are less likely to be seen in reality than the other. All input images were scaled to 60×60 pixels prior to the processing. The following classifier training strategies were compared with our approach:

(1) *PCA with histograms of oriented gradients (HOG)*: Each input image was divided into 6×6 adjacent regions. For each subregion a 6-bin histogram of gradient orientations was calculated and scaled to the range $[0, 1]$. Six-dimensional vectors yielded a 600-dimensional vector for each image after concatenation. PCA was employed to select a 16D linear combination of the original dimensions where 98% of the global data variance resided. Sign classification was done via nearest neighbour selection of the closest class mean in a sense of L2 metric.

(2) *AdaBoost with Haar wavelet features*: We used the approach of Jones and Viola [18] to learn a sign similarity measure from example image pairs. Resampling from the space of all possible negative example pairs was adopted to enable tractable training. Haar wavelet filters shown in Fig. 12 were used to construct weak classifiers within the AdaBoost framework. Size of each rectangular component of each filter satisfied $w, h = \{4, 8\}$ px and the filters were shifted by half of that size along each dimension, yielding a large over-complete space of input features. Such filters were computed independently in grey scale as well as in the images with red and blue colours enhanced, according to the first two transforms in Eq. (1). Classification of each

Table 4

Confusion matrices illustrating the classification accuracy obtained using different image representations and feature selection methods: (a) HOG/PCA, (b) Haar/AdaBoost, (c) HOG/AdaBoost, (d) CDT and class-specific discriminative local regions/forward search (our method).



Under each confusion matrix the total correct classification rate is shown.

test image was done by picking the label of the class prototype image for which the 100-feature boosted classifier yielded the maximum response when evaluating a pair comprised of this prototype and the image being tested. Prototype images were chosen randomly from the natural images available for each class.

(3) *AdaBoost with HOG features*: The same technique was used as in (2), but using HOGs as an underlying image representation. Histograms were computed in rectangular regions of size satisfying $w, h = \{6, 8, 10\}$ px, and shifted by half of the region size along each dimension.

Table 4 shows confusion matrices obtained for the test set using the abovementioned methods as well as our method involving CDT image representation and discriminative local region extraction. The results lead us to several conclusions. First, PCA does not seem to be an adequate technique for classifying similar object classes as it

merely captures the global variance of the data, but does not take class membership into account. On the other hand, a variant of AdaBoost introduced in [18] offers an elegant method of learning object dissimilarities and hence, unlike the standard AdaBoost, provides solution to any multi-class problem. However, the resulting classifier, despite being more complex, did not outperform the one obtained using our method.⁸

In general, the accuracy of the classifiers induced from the massive volumes of natural data does not seem to compensate the huge effort required to collect such data. This is particularly the case in

⁸ It should be noted that, in comparison to the dynamic recognition experiment, the significantly worse classification rates can be attributed to the lack of temporal integration.

our problem where the difficulty in obtaining a sufficient number of very rare signs' images necessitates reduction of the problem to the most popular classes. In this light, the proposed approach does not only offer a simpler training and a competitive recognition accuracy, but is also insensitive to the differences in the occurrence of the targeted signs in the real world. In addition, it seems to be more intuitive than the data-driven methods as the idealised templates provide sufficient knowledge about the unique appearance of traffic signs.

7. Conclusions

In this paper we have introduced a novel method for image representation and discriminative local feature selection, proving its usefulness in the task of traffic sign recognition. It was shown that on top of a discrete-colour image representation, a distance metric based on the *Colour Distance Transform*, and a forward feature selection technique, highly discriminative sign descriptors can be built from idealised templates based on the principle of *one-vs-all* dissimilarity maximisation. With these descriptors available, a conventional classifier can compete with the state-of-the-art methods, processing the input video sequences in close to real time. In comparison to the previous studies, our method seems attractive in the three aspects. First, feature extraction is performed directly from the publicly available template sign images, which makes the training effortless compared to the data-driven methods, such as AdaBoost. Second, each template is treated on an individual basis which is reflected in the number, position, and importance of the local image regions extracted in order to achieve a desired, globally set level of dissimilarity from the remaining templates. Finally, by using CDT we have shown that the proposed description of signs, although derived from the ideally clean template images, is suitable for modelling the intra-class appearance variability of traffic signs detected in the noisy traffic video.

References

- [1] R.E. Kalman, A new approach to linear filtering and prediction problems, *Transactions of the ASME Journal of Basic Engineering* 82 (1960) 35–45.
- [2] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society (B)* 39 (1) (1977) 138.
- [3] G. Borgefors, Distance transformations in digital images, *Computer Vision, Graphics, and Image Processing* 34 (3) (1986) 344–371.
- [4] D.M. Titterton, *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York, 1986.
- [5] D.B. Rubin, Using the SIR algorithm to simulate posterior distributions, *Bayesian Statistics* 3 (1988) 395–402.
- [6] P. Pudil, J. Novovicová, J. Kittler, Floating search methods in feature selection, *Pattern Recognition Letters* 15 (11) (1994) 1119–1125.
- [7] G. Piccoli, E. De Micheli, P. Parodi, M. Campani, A robust method for road sign detection and recognition, *Image and Vision Computing* 14 (3) (1996) 209–223.
- [8] Y. Aoyagi, T. Asakura, A study on traffic sign recognition in scene image using genetic algorithms and neural networks, in: *Proceedings of the 1996 IEEE IECON 22nd International Conference on Industrial Electronics, Control, and Instrumentation*, vol. 3, 1996, pp. 1838–1843.
- [9] A. de la Escalera, L.E. Moreno, M.A. Salichs, J.M. Armingol, Road traffic sign detection and classification, *IEEE Transactions on Industrial Electronics* 44 (6) (1997) 848–859.
- [10] D. Gavrilă, Multi-feature hierarchical template matching using distance transforms, in: *Proceedings of the IEEE International Conference on Pattern Recognition*, Brisbane, Australia, 1998, pp. 439–444.
- [11] M. Akmal Butt, P. Maragos, Optimum design of chamfer distance transforms, *IEEE Transactions on Image Processing* 7 (10) (1998) 1477–1484.
- [12] L. Zhang, F. Lin, B. Zhang, A CBIR method based on color-spatial feature, in: *Proceedings of the IEEE Region 10 Conference TENCEN 99*, vol. 1, 1999, pp. 166–169.
- [13] P. Paclík, J. Novovicová, P. Pudil, P. Somol, Road signs classification using the Laplace kernel classifier, *Pattern Recognition Letters* 21 (13–14) (2000) 1165–1173.
- [14] P. Douville, Real-time classification of traffic signs, *Real-Time Imaging* 6 (3) (2000) 185–193.
- [15] J. Miura, T. Kanda, Y. Shirai, An active vision system for real-time traffic sign recognition, in: *Proceedings of the IEEE Conference on Intelligent Transportation Systems*, Darborn, MI, USA, 2000, pp. 52–57.
- [16] T. Kadir, M. Brady, Scale, saliency and image description, *International Journal of Computer Vision* 45 (2) (2001) 83–105.
- [17] C.-Y. Fang, S.-W. Chen, C.-S. Fuh, Roadsign detection and tracking, *IEEE Transactions on Vehicular Technology* 52 (5) (2003) 1329–1341.
- [18] M. Jones, P. Viola, Face recognition using boosted local features. Technical Report TR2003-25, 2003.
- [19] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* 60 (2) (2004) 91–110.
- [20] H. Fleyeh, Color detection and segmentation for road and traffic signs, in: *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems*, Singapore, vol. 2, 2004, pp. 809–814.
- [21] P. Viola, M. Jones, Robust real-time object detection, *International Journal of Computer Vision* 57 (2) (2004) 137–154.
- [22] A. de la Escalera, J.M. Armingol, J.M. Pastor, F.J. Rodríguez, Visual sign information extraction and identification by deformable models for intelligent vehicles, *IEEE Transactions on Intelligent Transportation Systems* 5 (2) (2004) 57–68.
- [23] G. Loy, N. Barnes, D. Shaw, A. Robles-Kelly, Regular polygon detection, in: *Proceedings of the 10th IEEE International Conference on Computer Vision*, vol. 1, 2005, pp. 778–785.
- [24] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, T. Koehler, A system for traffic sign detection, tracking and recognition using color, shape, and motion information, in: *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2005, pp. 255–260.
- [25] X.W. Gao, L. Podladchikova, D. Shaposhnikov, K. Hong, N. Shevtsova, Recognition of traffic signs based on their colour and shape features extracted using human vision models, *Journal of Visual Communication and Image Representation* 17 (4) (2006) 675–685.
- [26] M.A. Garcia-Garrido, M.A. Sotelo, E. Martin-Gorostiza, Fast traffic sign detection and recognition under changing lighting conditions, in: *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2006, pp. 811–816.
- [27] P. Paclík, J. Novovicová, R.P.W. Duin, Building road-sign classifiers using a trainable similarity measure, *IEEE Transactions on Intelligent Transportation Systems* 7 (3) (2006) 309–321.
- [28] A. Ruta, Y. Li, X. Liu, Towards real-time traffic sign recognition by class-specific discriminative features, in: *Proceedings of the 18th British Machine Vision Conference*, Coventry, United Kingdom, vol. 1, 2007, pp. 399–408.
- [29] A. Ruta, Y. Li, X. Liu, Traffic sign recognition using discriminative local features, in: *Proceedings of the 7th International Symposium on Intelligent Data Analysis*, Ljubljana, Slovenia, 2007, pp. 355–366.
- [30] B. Mecánico, Road signs arranged by country. [www] Available from: (<http://www.elve.net/rcoulst.htm>) [Accessed September 22, 2008].

About the Author—ANDRZEJ RUTA received his M.Sc. degree in computer science from the AGH University of Science and Technology, Krakow, Poland, in 2005. Currently he is a PhD student working in the Centre for Intelligent Data Analysis within the School of Information Systems, Computing and Mathematics at Brunel University, West London, UK. His present research is focused on machine learning, visual object recognition, and realtime traffic video analysis. Other research interests include data mining, pattern recognition, nature-inspired information systems, and financial time series prediction.

About the Author—YONGMIN LI is a senior lecturer in the School of Information Systems, Computing and Mathematics at Brunel University, West London, UK. He received his M.Eng. and B.Eng. in control engineering from Tsinghua University, China, in 1990 and 1992, respectively, and Ph.D. in computer vision from Queen Mary, University of London in 2001. Between 2001 and 2003 he worked as a research scientist in the Content and Coding Lab at BT Exact (formerly the British Telecom Laboratories). In 2001 he was the winner of the British Machine Vision Association (BMVA) best scientific paper award and the winner of the best paper prize of the 2001 IEEE International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems. His current research interests cover the areas of computer vision, image processing, video analysis, machine learning, and pattern recognition.

About the Author—XIAOHUI LIU is a professor of Computing in the School of Information Systems, Computing and Mathematics at Brunel University, West London, UK, where he leads the Centre for Intelligent Data Analysis, a centre of excellence for multidisciplinary research involving artificial intelligence, dynamic systems, signal processing and statistics, particularly for biomedical and engineering applications. He serves on the editorial boards of four computing journals, founded the biennial international conference series on IDA in 1995, and has given invited or keynote talks at a number of computing, bioinformatics, and statistics conferences. His research interests include bioinformatics, data mining, intelligent systems, medical informatics, optimisation, signal and image processing, time series, and visualisation.