IDIAP RESEARCH REPORT

Confidence-based Cue Integration for Visual Place Recognition

Andrzej Pronobis [a] Barbara Caputo [b] [c]

IDIAP–RR 07-17

April 2007

submitted for publication

[a] NADA/CAS, KTH, Sweden -pronobis@nada.kth.se
[b] IDIAP - bcaputo@idiap.ch
[c] Ecole Polytechnique Fédérale de Lausanne (EPFL) - 1015 Lausanne (Switzerland)

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07-17

Confidence-based Cue Integration for Visual Place Recognition

Andrzej Pronobis Barbara Caputo

April 2007

submitted for publication

Abstract. A distinctive feature of intelligent systems is their capability to analyze their level of expertise for a give task; in other words, they know what they know. As a way towards this ambitious goal, this paper presents an algorithm for recognition able to measure its own level of confidence and, in case of uncertainty, to seek for extra information so to increase its own knowledge and ultimately achieve better performance. We focus on the visual place recognition problem for topological localization, and we take an SVM approach. We propose a new method for measuring the confidence level of the classification output, based on the distance of a test image and the average distance of training vectors. This method is combined with a discriminative accumulation scheme for cue integration. We show with extensive experiments that the resulting algorithm achieves better performances for two visual cues than the classic single cue SVM on the same task, while minimising the computational load. More important, our method provides a reliable measure of the level of confidence of the decision.

1 Introduction

A key competence for an autonomous agent is the ability to localize itself in the world. Vision-based localization represents a challenge for the research community, because the visual information tends to be noisy and difficult to analyze. Still, this research line is attracting more and more attention, and several methods have been proposed using vision alone [1, 2, 3], or combined with more traditional range sensors [4, 5]. The increasing activity in this research area comes firstly from the portability and cost-effectiveness of visual sensors; secondly, from the specific type of information that only these sensors can bring. This is the case for instance in place categorization or understanding, where the semantic information plays a crucial role. Furthermore, visual place recognition can be applied as a method for loop closing, scalability issues, or recovery from the kidnapped robot problem.

A vast majority of algorithms proposed so far were designed to provide as output a hard decision: the system is trained to recognize a fixed and pre-defined set of environments (e.g. kitchen, corridor, office etc.) and then, when presented with a test image, it classifies it as one of the possible places, but little or nothing is generally said regarding the *confidence* of this decision or other possible hypotheses. Measuring confidence, or knowing what is known, is a fundamental concept for autonomous robots. Indeed, in many real-world applications it is more desirable to abstain from action because of a self-recognized lack of confidence, rather than take a hard decision which might result in a costly error. Thus, introducing a confidence measure in a recognition algorithm allows to provide reliability despite constrained performance of the algorithm or lack of updated information, and makes it possible to evaluate when it is necessary to seek for extra information (e.g. from multiple cues or modalities) in order to achieve a confident decision.

It is possible to define a confidence measure for any pattern recognition algorithm: for probabilistic methods, it will be related to the posterior probability of the image at hand; for discriminative classifiers, it will be related to the distance from the separating hyperplane. In this paper, we will focus on large margin classifiers, specifically on Support Vector Machines (SVMs), even if most of the concepts and ideas we will propose can be easily extended to any other margin-based discriminative methods. We build on our previous work on place recognition, where we presented an SVM-based method able to recognize indoor environments under severe illumination changes and across a time span of several weeks [3]. Our first contribution is the introduction of a method for ranking the hypotheses generated by the classifier and measuring their confidence. The method is based on the distance from the hyperplane and the average distance of each training class. We present experiments showing that our confidence measure gives a better performance compared with the classic hard decision SVM and, more important, a decision that is more informative of the level of knowledge of the robot.

Once a system is able to output not only its guess, but also the level of confidence of the guess, action should be taken. Indeed, we can expect that when an image is classified with a low level of confidence, it is because the algorithm doesn't have enough information. A possible way to increase the knowledge, and thus the confidence, is to use additional information such as both global and local visual descriptors, or laser-based geometrical data, and combine them through an integration scheme. An effective method for visual cue integration using SVMs has been proposed in [6], called Discriminative Accumulation Scheme (DAS). A second contribution of this paper is to apply that algorithm to the domain of vision for robotics. We also propose its generalized version (Generalized DAS), that can be built on top of our confidence estimation method. Experiments confirm the effectiveness of the approach and show that G-DAS consistently outperforms the original DAS.

While using multiple visual cues improves both classification accuracy and relative confidence, it is computationally expensive (more features to compute and classify), which is undesirable for an autonomous agent. Ideally, a system should use additional information only when necessary, i.e. only when the level of confidence of a single cue is not such to obtain a reliable decision. The final contribution of this paper is to combine the G-DAS framework with the confidence estimation approach, so that multiple cues are used only when they are necessary to disambiguate low-confidence cases. Our experiments on both local and global visual cues show that the proposed approach reduces the computational load of about 55% in average, while achieving the same performance obtained by using G-DAS on all the images.

The rest of the paper is organized as follows: after an overview of previous work on confidence measures and cue integration (Section 2), Section 3 gives a brief description of the methodology used further. Section 4 describes our confidence estimation method and evaluates its effectiveness for visual place recognition. Section 5 reviews DAS, presents our generalized version of the algorithm and assesses its performance; Section 6 shows how by combining the two techniques we achieve a better overall performace while reducing the computational load. The paper concludes with a summary and possible avenues for future research.

2 Related Work

We are not aware of confidence estimation and/or cue integration methods within the robotics literature for visual place recognition. However, computing confidence estimates for discriminative classifiers is an open problem in machine learning. Although classifiers like K-NN, ANN, or SVM output numeric scores for class membership, some experiments show that, when used directly, they are not well correlated with classification confidence [7]. Several authors attacked this problem by developing more sophisticated measures such as probability estimates obtained by trained sigmoid function [8] with extensions for multi-class problems [9], or relative distance from the separating hyperplane, normalized with the average class distance from the plane [10]. More comments on their performance can be found in Section 4.

Visual cue integration via accumulation was first proposed in a probabilistic framework by Poggio *et al.*[11], and then further explored by Aloimonos and Shulman [12]. The idea was then extended to SVMs by Nilsback and Caputo [6] (DAS). The resulting method showed remarkable performances on object recognition applications and together with its generalized version (G-DAS) is used here as a cue integration scheme for disambiguating classes with low confidence estimate.

3 A Few Landmarks

This section serves as a base for the results and theory presented further. We present the common scenario and methodology used during all experimental evaluations (Sections 3.1 and 3.2), we briefly review SVMs (Section 3.4) and the visual descriptors used throughout the paper (Section 3.3).

3.1 Experimental Scenario

The algorithms presented in this paper have been tested in the domain of mobile robot topological localization. As benchmarking data we used the IDOL (Image Database for rObot Localization [13]) database which was introduced in [3] in order to test robustness of our discriminative approach to visual place recognition in real-world scenario and under varying illumination conditions. The database comprises sequences of images of places acquired using cameras of resolution 320x240 pixels mounted at different heights (98cm and 36cm) on two mobile robot platforms, the PeopleBot Minnie and the PowerBot Dumbo. The acquisition was performed in a five room subsection of a larger office environment, selected in such way that each of the five rooms represented a different functional area: a one-person office, a two-persons office, a kitchen, a corridor, and a printer area (part of the corridor). Example pictures showing interiors of the rooms are presented in Fig. 1.

The appearance of the rooms was captured under three different illumination conditions: in cloudy weather, in sunny weather, and at night. The robots were manually driven through each of the five rooms while continuously acquiring images at the rate of 5 frames per second. Each image was then labelled as belonging to one of the rooms according to the position of the robot estimated using a laser-based localization method. The acquisition was performed twice for each robot and illumination condition, resulting in 12 image sequences in total over a span of time of more than two weeks. In consequence, the sequences captured variability introduced not only by illumination but also natural activity in the environment (presence/absence of people, furniture relocated etc.). Example images illustrating the captured variability for both robot platforms are shown in Fig. 1.

3.2 Experimental Procedure

As a basis for the experiments, we used the place recognition system presented in [3], which is built around Support Vector Machines [14], and a rich global descriptor [15]. While designing the system, we followed the assumption that global configuration of a scene is informative enough for recognition and obtained good performance despite variations captured in the IDOL database. In this work, in order to increase robustness, we additionally used the SIFT descriptor [16] that have already been proved successful in the domain of vision-based localization [1].

Following [3], we took a fully supervised approach and assumed that during training each room is represented by a collection of images capturing its visual appearance under various viewpoints, at fixed time and illumination setting. During testing, the algorithm is presented with images of the same rooms, acquired under roughly similar viewpoints but possibly under different illumination conditions, and after some time. The goal is to recognize each single image seen by the system. As in [3], we considered three sets of experiments for three types of problems of different complexity. In case of each single experiment, training was always performed on one image sequence subsampled to 1 fps (every fifth image), and testing was done using a full sequence. The first set consisted of 12 experiments performed on different combinations of training and test data, acquired using the same robot platform and under similar illumination conditions. For the second set of experiments, we used 24 pairs of sequences captured under different illumination conditions. Finally, the third set was performed on 24 pairs of training and test sequences acquired under similar illumination settings but using a different robot. As a measure of performance we used the percentage of properly classified images calculated separately for each of the rooms and then averaged with equal weights independently of the number of images acquired in each room.

3.3 Image Representations

In this work, we employed two types of visual cues, global and local, extracted from the same image frame. As global representation we used the Composed Receptive Field Histograms (CRFH) [15], a multi-dimensional statistical representation of responses of several image filters. Computational costs were reduced by using a sparse and ordered histogram representation, as proposed in [15]. Following [3], we used histograms of 6 dimensions, with 28 bins per dimension, computed from second order normalized Gaussian derivative filters applied to the illumination channel at two scales.

We used the SIFT descriptor [16] in order to obtain the local image representation. SIFT represents the local image patches around interest points characterized by coordinates in the scalespace in the form of histograms of gradient directions. In order to find the coordinates of the interest points, we used the Harris-Laplacian detector [17], a scale invariant extension of the Harris corner detector.

3.4 Support Vector Machines

Consider the problem of separating the set of training data $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ into two classes, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label. If we assume that the two classes can be separated by a hyperplane in some Hilbert space $\mathcal{H}$, then the optimal separating hyperplane is the one which has maximum distance to the closest points in the training set resulting in a discriminant function

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{1}$$

The classification result is then given by the sign of $f(\boldsymbol{x})$. The values of α_i and b are found by solving a constrained minimization problem, which can be done efficiently using the SMO algorithm [14]. Most of the α_i's take the value of zero; those $\boldsymbol{x}_i$ with nonzero α_i are the "support vectors". In case where the two classes are non-separable, the optimization is formulated in such way that the classification error is minimized and the final solution remains identical.

The mapping between the input space and the usually high dimensional feature space $\mathcal{H}$ is done using the kernel function $K(\boldsymbol{x}_i, \boldsymbol{x})$. Several kernel functions have been proposed for visual applications; in this paper we will use the χ^2 kernel [18] for the global CRFH descriptors, and the match kernel proposed in [19] for the local SIFT descriptors. Both have been used in our previous work on SVM-based place recognition, obtaining good performances.

The extension of SVM to multi class problems can be done mainly in two ways:

1. *One-against-All (OaA) strategy.* If M is the number of classes, M SVMs are trained, each separating a single class from all remaining classes. The decision is then based on the distance of the classified sample to each hyperplane. Typically algebraic distance ($f(\boldsymbol{x})$) is used and the final output is the class corresponding to the hyperplane for which the distance is largest.
2. *One-against-One (OaO) strategy.* In this case, $M(M-1)/2$ two-class machines are trained for each pair of classes. The final decision can then be taken in different ways, based on the $M(M-1)/2$ outputs. A popular choice is to consider as output of each classifier the class label and count votes for each class; the test image is then assigned to the class that received more votes. Another alternative is to use signed distance from the hyperplane and sum distances for each class. Other solutions based on the idea to arrange the pairwise classifiers in trees, where each tree node represents an SVM, have also been proposed [20, 14]. In this paper, we will use the voting-based method, which we found to constantly outperform the second alternative in our preliminary experiments.

4 Confidence Estimation

This section presents our approach to the problem of ranking hypotheses generated by the classifier and measuring their confidence. We first describe two methods based on the standard OaO and OaA multi-class extensions and our modified version of the OaA principle. Then, we show benchmark experiments evaluating the performance of the methods on visual data. The algorithms presented here will be one of the building block of the confidence-based cue integration scheme we will introduce in Section 6.

4.1 The algorithms

As already mentioned, discriminative classifiers do not provide any out-of-the-box solution for estimating confidence of the decision; however, it is possible to derive confidence information and hypotheses ranking from the distances between the samples and the hyperplanes. In case of SVMs, this can be done very efficiently thanks to the use of kernel functions and does not require additional processing in the training phase (as opposed to probability estimation methods like [8]). As it will be shown by experiments, despite its simplicity, such approach can yield good results when applied to complex problems such as visual place recognition. We stress that, since it is based on the generated hyperplanes, performance will depend on how well the model reflects the statistics of the test data. In other words, the approach will work best for cases where the difficulty comes from the inability to perfectly separate the training samples and still provide good generalization capabilities.

It is straightforward to extend the standard OaO and OaA multi-class methods so that additional information about the decision becomes available. Let us present the methods using a more general notation and introduce a variable $V_h(\boldsymbol{x})$, which will be a distance-based score assigned by the hyperplane h to the sample $\boldsymbol{x}$. In case of the two standard algorithms, the score will just be equal to the distance of the test sample to the hyperplane: $V_h(\boldsymbol{x}) = D_h(\boldsymbol{x})$. Typically, the value of the discriminant function is used as a distance measure ($D_h(\boldsymbol{x}) = f_h(\boldsymbol{x})$). In order to find the best hypothesis j^*, we follow the rules described in Section 3.4:

- *for the OaO strategy:*

$$j^* = \underset{j=1\ldots M}{\operatorname{argmax}} \left|\{i : i \in \{1 \ldots M\},\ i \neq j,\ V_{i,j}(\boldsymbol{x}) > 0\}\right|,$$

  where the indices i, j are used to denote the hyperplane separating class i from class j.

- *for the OaA strategy:*

$$j^* = \underset{j=1\ldots M}{\operatorname{argmax}} \{V_j(\boldsymbol{x})\},$$

  where V_j is the score assigned by the hyperplane separating class j from the other classes.

If now we think of the confidence as a measure of unambiguity of the decision, we can define it as:

- *for OaO,* the minimal score (distance) to the hyperplanes separating the first hypothesis and the other classes:

$$C(\boldsymbol{x}) = \min_{j=1\ldots M,\ j\neq j^*} \{V_{j^*,j}(\boldsymbol{x})\}$$

- *for OaA,* the difference between the maximal and the next largest score:

$$C(\boldsymbol{x}) = V_{j^*}(\boldsymbol{x}) - \max_{j=1\ldots M,\ j\neq j^*} \{V_j(\boldsymbol{x})\}$$

The value $C(\boldsymbol{x})$ can be thresholded for obtaining a binary confidence information. Confidence is then assumed if $C(\boldsymbol{x}) > \tau$ for threshold τ. The values $V_{j^*,j}(\boldsymbol{x})$ (for OaO) and $V_j(\boldsymbol{x})$ (for OaA) can also be used to rank the hypotheses and find between which of them the classifier is uncertain.

The outcome of the algorithms described above depends only on the distances of the test sample to the hyperplanes, that for SVMs is determined by the vectors lying close to the class boundaries (the support vectors). To make it more dependent on the distribution of all available training data, we suggest to use the OaA principle and redefine the score $V_j(\boldsymbol{x})$ to be equal to the distance from the average distance of the training samples to the hyperplane (see Fig. 2 for an illustration):

$$V_j(\boldsymbol{x}) = \left|\widehat{D}_j - D_j(\boldsymbol{x})\right|.$$

Thus, we do not measure how far the test sample is from the hyperplane, but how close it is to the training data belonging to one of the classes. The best hypothesis can be determined by the following rule:

$$j^* = \underset{j=1\ldots M}{\operatorname{argmin}} \{V_j(\boldsymbol{x})\}. \tag{2}$$

Using the same definition of confidence as above, we get:

$$C(\boldsymbol{x}) = \min_{j=1\ldots M,\ j\neq j^*} \{V_j(\boldsymbol{x})\} - V_{j^*}(\boldsymbol{x}). \tag{3}$$

As in case of the previous algorithms, we can order the hypotheses using the values of $V_j(\boldsymbol{x})$ and obtain hard confidence information by thresholding. An explanation on a real example from one of our experiments is shown in Fig. 3.

4.2 Experimental Evaluation

We preformed a benchmark evaluation of the three confidence estimation methods (the two methods based on the standard OaO and OaA multi-class extensions and the method based on the new modified version of OaA) on the IDOL database. As described in Section 3.2, we performed three sets of experiments: training and testing under stable illumination conditions, varying illumination conditions, and recognition across different robotic platforms. For all experiments we measured the performance of the algorithms for a range of values of the confidence threshold. We used two measures of performance in order to analyse different properties of the methods. First, for each value of the confidence threshold, we calculated the classification rate (percentage of properly classified test images) only for those test samples for which the decision was regarded as confident. As a second measure we used the classification rate calculated for all samples and including additional hypotheses between which the algorithm was unsure when the confidence was below the threshold. For example, if for a given value of the threshold the decision was "kitchen or corridor" and the test image was acquired in one of these rooms, the decision was counted as correct.

The average results obtained for the global features (CRFH) are presented in Fig. 4. The experiments were repeated also for local features (SIFT); however, the results showed the same trends and thus are omitted for space reasons [1]. To obtain these results, we used the value of the discriminant function as a distance measure ($D_h(\boldsymbol{x}) = f_h(\boldsymbol{x})$). We performed identical experiments for two other distance measures: the distance of a sample to its normal projection onto the hyperplane and relative distance normalized by the average class distance to the plane [10]; however, the results clearly showed the advantage of the solutions based on the value of $f(\boldsymbol{x})$.

The plots presented in Fig. 4 show the dependency between the classification rates and the percentage of images of the test sequence for which the classifier was not confident, given some value of threshold. The classification rates for hard-decision SVM are marked on the vertical axis (initial values, all decisions treated as confident). It can be observed that the classification rate calculated for the confident decisions only (Fig. 4a) is increasing for all methods as the percentage of unsure decisions grows. This means that the algorithms tend to eliminate the misclassified samples. It is clear that the modified OaA approach performs best with respect to this measure. Moreover, we can see that the method consistently delivers best classification rates when hard decisions are considered. The advantage in terms of classification rate varyes from +0.4% to +3.5% with respect to standard OaA and +1.5% to +5.7% with respect to standard OaO and grows with the complexity of the problem.

Additional conclusions can be drawn from the analysis of the second performance measure (Fig. 4b). First, we see that if we tolerate soft decisions in e.g. 30% of cases, the resulting classification rate increases from +5.2% (Fig. 4b, left) to +12% (Fig. 4b, right) and can even reach 99% in case of experiments performed for similar illumination conditions. Second, it is still visible that both OaA-based methods consistently outperform the OaO-based algorithm, and the modified version of the OaA strategy in general achieves the best performance. This time, however, the advantage of the modified OaA with respect to the algorithm based on the standard approach is smaller and decreases as the number of unsure decisions grows. This makes us conclude that the modified OaA method is better when it comes to finding and estimating confidence of the best hypothesis. However, the standard OaA-based algorithm is similarly or even more (Fig. 4b, right) efficient for ranking hypotheses. This property of the modified algorithm may become important if additional information could be used to improve classification results for the decisions in cases when the classifier is not confident enough.

5 Cue Integration

Last section showed the importance of defining an effective confidence measures for SVM, and its impact on classification accuracy. Still, once the algorithm is able to measure an unsatisfactory level of confidence, it should react accordingly. The most desirable action would of course lead to higher confidence and accurate classification; one of the possible way to achieve this result is to use multiple cues and combine them effectively. In this section, we introduce a generalization of the integration scheme proposed in [6] to a wider class of multi-class extensions, and we present experimental evidence of its efficiency. How to combine these cue integration schemes with confidence-based classification approach will be the subject of Section 6.

[1] Classification rates for local features and hard-decision SVMs can be found in Section 5

5.1 Generalized Discriminative Accumulation Scheme

Suppose we are given M visual classes and, for each class, a set of n_j training images $\{\boldsymbol{I}_i^j\}_{i=1}^{n_j}$, $j = 1, \ldots, M$. Suppose also that, from each image, we extract a set of P different cues $\{T_p(\boldsymbol{I}_i^j)\}_{p=1}^{P}$ (the cues could also be different modalities). The goal is to perform recognition using all the cues. The original DAS algorithm consists of two steps:

1. *Single-cue SVMs.* From the original training set $\{\{\boldsymbol{I}_i^j\}_{i=1}^{n_j}\}_{j=1}^{M}$, containing images belonging to all M classes, define P new training sets $\{\{T_p(\boldsymbol{I}_i^j)\}_{i=1}^{n_j}\}_{j=1}^{M}$, $p = 1, \ldots, P$, each relative to a single cue. For each new training set train a multi-class SVM. In general, kernel functions may differ from cue to cue. Model parameters can be estimated during the training step via cross validation. In case of the original DAS algorithm, the standard OaA multi-class extension was used. Then, given a test image $\boldsymbol{I}$, for each single-cue SVM the algebraic distance to each hyperplane $f_j^p(T_p(\boldsymbol{I}))$, $j = 1, \ldots, M$ was computed according to Eq. 1.

2. *Discriminative Accumulation.* After all the distances were collected $\{f_j^p\}_{p=1}^{P}$, for all the M hyperplanes and the P cues, the image $\boldsymbol{I}$ was classified using their linear combination:

$$j^* = \underset{j=1}{\overset{M}{\operatorname{argmax}}} \left\{ \sum_{p=1}^{P} a_p f_j^p(T_p(\boldsymbol{I})) \right\}, \ a_p \in \Re^+.$$

   The coefficients $\{a_p\}_{p=1}^{P}$ can also evaluated via cross validation during the training step.

The original algorithm performed accumulation at the level of algebraic distances from the hyperplanes f_j^p, obtained from a standard OaA multi-class SVM. As shown in Section 4, there are other methods available, and it is possible to introduce more effective multi-class algorithms based on the OaA principle. We thus propose to extend the DAS framework to be applicable also for the other methods; we call this extension the Generalized Discriminative Accumulation Scheme (G-DAS). The discriminative accumulation is here performed at the level of the scores V_h (see Section 4):

$$V_h^{\Sigma P}(\boldsymbol{I}) = \sum_{p=1}^{P} a_p V_h^p(T_p(\boldsymbol{I})), \ a_p \in \Re^+. \tag{4}$$

As a result, any multi-class extension can be used within the G-DAS framework (both OaA and OaO based) in order to obtain the final decision.

5.2 Experimental Evaluation

We evaluated the effectiveness G-DAS for the visual place recognition problem by running the three series of experiments described in Section 3.2. SIFT and CRFH were used as features, χ^2 and match kernel as similarity measures for the nonlinear SVMs, and kernel parameters as well as weighting coefficients for the accumulation schemes were determined via cross validation.

Fig. 5 shows the recognition results obtained using a single cue SVM, with global or local descriptors, and those obtained using the G-DAS algorithm. For all those three different approaches, we used three different multi-class extensions: standard OaO, standard OaA and our new modified OaA. Note that, when using standard OaA, G-DAS corresponds to the original DAS algorithm.

A first comment is that for all three different multi-class extensions, for all the three series of experiments, the accumulation scheme clearly achieves consistently better results than the single cues approaches. The gain in performance goes from a minimum of +1.9% in accuracy, obtained for the stable illumination condition experiments (Fig. 5a) to a maximum of +7.8%, obtained for the varying illumination (Fig. 5b), with respect to the CRFH only, using the modified OaA approach. The increase in performance grows with the difficulty of the task and is on average a +2% for stable illumination, +5% for varying illumination and +6% for recognition across platforms. A second comment is that G-DAS with our modified OaA consistently performs better than the original DAS, for all the three scenarios; this confirms the effectiveness of this confidence measure for visual recognition. An important property of the DAS algorithm, which is also preserved by G-DAS, is the ability to classify correctly images even when each of the single cues used gives misleading information. Fig. 6 shows an example of this behavior: the test image is misclassified as 'one-person office' by using CRFH, and as 'corridor' by using SIFT; by combining these two cues in G-DAS, the image is correctly classified as 'two-persons office'. We can then conclude that G-DAS is an effective method for cue integration for visual place recognition in realistic settings.

6 Confidence-based Cue Integration

As it was motivated in Section 5, a desirable behaviour of a system aware of its own ignorance would be to search for additional sources of information in order to achieve higher confidence. The presented results show that G-DAS can be effectively used for visual cue integration; however, it requires that both cues were available and used for classification even in cases when one cue is sufficient to obtain correct result, and the additional computational effort could be avoided. In this section, we present and experimentally evaluate a strategy allowing to greatly decrease the computational load and still maintain the high level of accuracy provided by multiple cues. We propose to employ the G-DAS framework for cue accumulation and extract the additional information only in cases when the confidence of a decision based on the cues available so far is not satisfactory.

6.1 The method

It is reasonable to assume that the confidence estimation methods presented in Section 4 can be used as efficient filters, filtering out the images for which G-DAS would be either not required or not effective. First, the experiments reported in Section 4 proved that the confidence estimation methods are able to eliminate the incorrect decisions. Second, both methods and the G-DAS framework operate on the distances calculated in the high dimensional feature space, and G-DAS is expected to be most effective in cases of low confidence (see the example in Fig. 6).

Suppose again that we are able to extract P different cues $\{T_p(\boldsymbol{I})\}_{p=1}^{P}$ from the input image $\boldsymbol{I}$. Let us assume that the cues are ordered. The order of the cues can be motivated by the computational cost associated with feature extraction and classification. To obtain the final decision we use the following algorithm:

1. Set $k = 1$.
2. Extract features for the k^{th} cue ($T_k(\boldsymbol{I})$).
3. Perform classification for the k^{th} cue and obtain the scores $V_h^k(T_k(\boldsymbol{I}))$ for all hyperplanes.
4. Perform cue integration for the cues $1 \dots k$ according to Eq. 4 and obtain the accumulative scores $V_h^{\Sigma k}(\boldsymbol{I})$.
5. Find the best hypothesis j_k^* and confidence estimates $C_k(I)$ based on the scores $V_h^{\Sigma k}(\boldsymbol{I})$.
6. If the confidence is below the threshold ($C_k(I) < \tau$) and $k < P$, increment k and go to step 2. Otherwise, use the obtained hypothesis as final decision ($j^* = j_k^*$).

The threshold τ is a parameter of the algorithm and allows to trade the accuracy for computational cost.

6.2 Experimental Evaluation

We performed an experimental evaluation of the confidence-based cue integration strategy for the two global (CRFH) and local (SIFT) visual cues. We tested solutions based on both CRFH and SIFT used as a primary cue. The experiments showed the advantage of the CRFH-based solution in terms of the number of images for which both cues had to be used to obtain accuracy identical with the one offered by G-DAS. Moreover, the local features are much more computationally expensive mainly due to matching process performed during classification.

In this paper, we report results for CRFH used as a primary cue. The plots presented in Fig. 7 clearly show that in order to obtain accuracy comparable with the one delivered by G-DAS used for all test images, it is necessary to use the second cue only in approximately 40% of cases. This is for the modified OaA multi-class extension, which once more outperformed the other confidence estimation methods. As already mentioned, in our case, feature extraction and classification was more costly for the local cue. As a result, the strategy presented here allowed to reduce the amount of computations by about 55% in average compared to G-DAS. Since the dependency between the number of images for which the second cue is used and the classification rate is highly non-linear, it can be advantageous to trade the accuracy for computational cost; e.g. to achieve gain of 70% of the one provided by G-DAS, the second cue should be used in 7% (stable illumination conditions, Fig. 7a) to 22% (varying illumination conditions, Fig. 7b) of cases only. Concluding, the power of multiple cues can be achieved for much lower computational cost, if information about the classifier's confidence is exploited.

7 Summary and Conclusion

This paper presented an effective approach to the problems of confidence estimation and cue integration for large-margin discriminative classifiers. We showed by extensive experiments, on problems of different complexity from the domain of visual place recognition, that exploiting available confidence information encoded in the classifier's outputs can greatly increase reliability of a system. When combined with a cue integration scheme, this results in a significantly increased performance for a relatively low computational cost. We used SVMs and combined local and global cues extracted from the same visual stimuli; all the presented methods could easily be extended to other large margin classifiers and to multiple modalities.

The potential of this approach can be used in many ways. Firstly, we plan to incorporate confidence information into an incremental learning framework and use it to trigger the learning procedure. Secondly, we want to create an active system able to autonomously search for cues in order to obtain confident decision. Finally, we will test our method in a multi-modal system and for integration of a larger number of cues.

References

[1] S. Se, D. Lowe, and J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *Proc. ICRA'01*.

[2] I. Ulrich and I. R. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proc. ICRA'00*.

[3] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen, "A discriminative approach to robust visual place recognition," in *Proc. IROS'06*.

[4] D. Kortenkamp and T. Weymouth, "Top. mapping for mobile robots using a combination of sonar and vision sensing," in *Proc. AAAI'94*.

[5] A. Tapus and R. Siegwart, "Incremental robot mapping with fingerprints of places," in *Proc. IROS'05*.

[6] M. E. Nilsback and B. Caputo, "Cue integration through discriminative accumulation," in *Proc. CVPR'04*.

[7] S. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh, "Generating est. of class. conf. for a case-based spam filter," in *Proc. ICCBR'05*.

[8] J. Platt, "Probabilistic outputs for SVMs and comparisons to regularized likelihood methods," in *Adv. in Large Margin Classifiers*, 2000.

[9] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class class. by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, 2004.

[10] J.-J. Kim, B.-W. Hwang, and S.-W. Lee, "Retrieval of the top n matches with support vector machines," in *Proc. ICPR'00*.

[11] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, 1985.

[12] J. Aloimonos and D. Shulman, *Integration of Visual Modules: an Extension of the Marr Paradigm.* Academic Press, 1989.

[13] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "The IDOL2 database," KTH, CAS/CVAP, Tech. Rep., 2006, Available at `http://cogvis.nada.kth.se/IDOL/`.

[14] N. Cristianini and J. S. Taylor, *An Introduction to SVMs and Other Kernel-based Learning Methods.* Cambridge University Press, 2000.

[15] O. Linde and T. Lindeberg, "Object recognition using composed receptive field histograms of higher dimensionality," in *Proc. ICPR'04*.

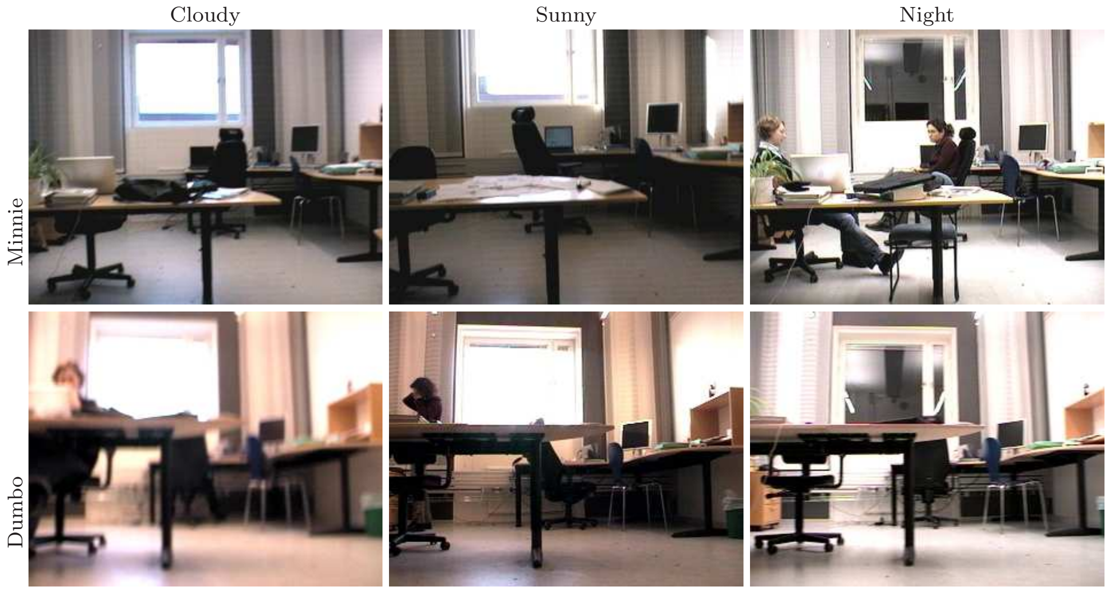
[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, 2004.

[17] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *Proc. ICCV'01*.

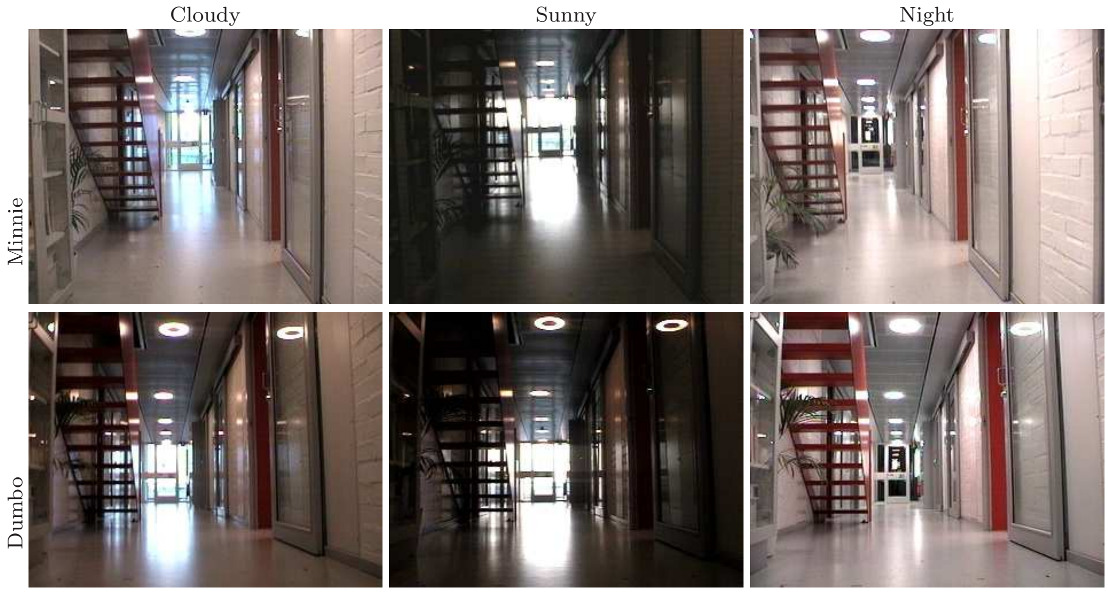
[18] O. Chapelle, P. Haffner, and V. Vapnik, "SVMs for histogram-based image classification," *IEEE Trans. Neur. Netw.*, vol. 10, no. 5, 1999.

[19] C. Wallraven, B. Caputo, and A. Graf, "Recognition with local features: the kernel recipe," in *Proc. ICCV'03*.

[20] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," in *Proc. NIPS'00*.



(a) Two-persons office



(b) Corridor

(c) Remaining rooms (Minnie at night)

Figure 1: Example pictures taken from the IDOL database showing the interiors of the rooms, variations occurring across platforms, as well as introduced by illumination changes and natural activity in the environment.

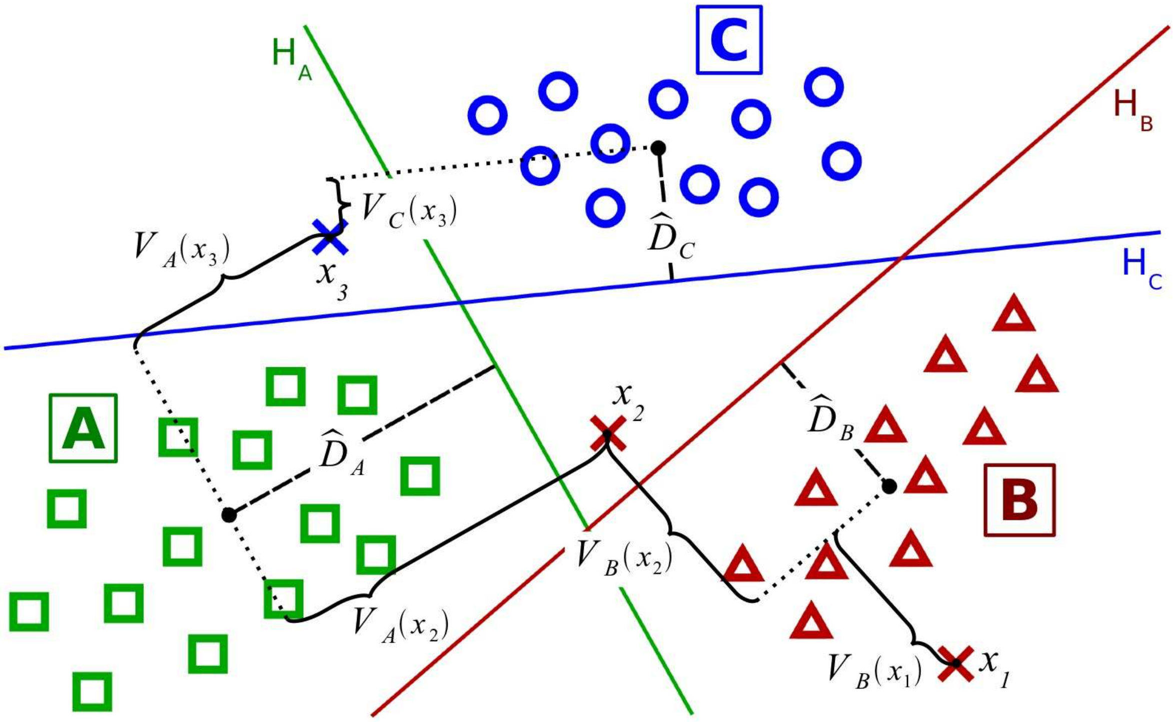


Figure 2: Artificial classification problem illustrating the way the scores $V_j(\boldsymbol{x})$ are calculated for classified samples in case of the modified OaA approach. We can observe that although the points $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are located approximately in the same distance from two hyperplanes, they are classified as belonging to class B and C respectively with high confidence.

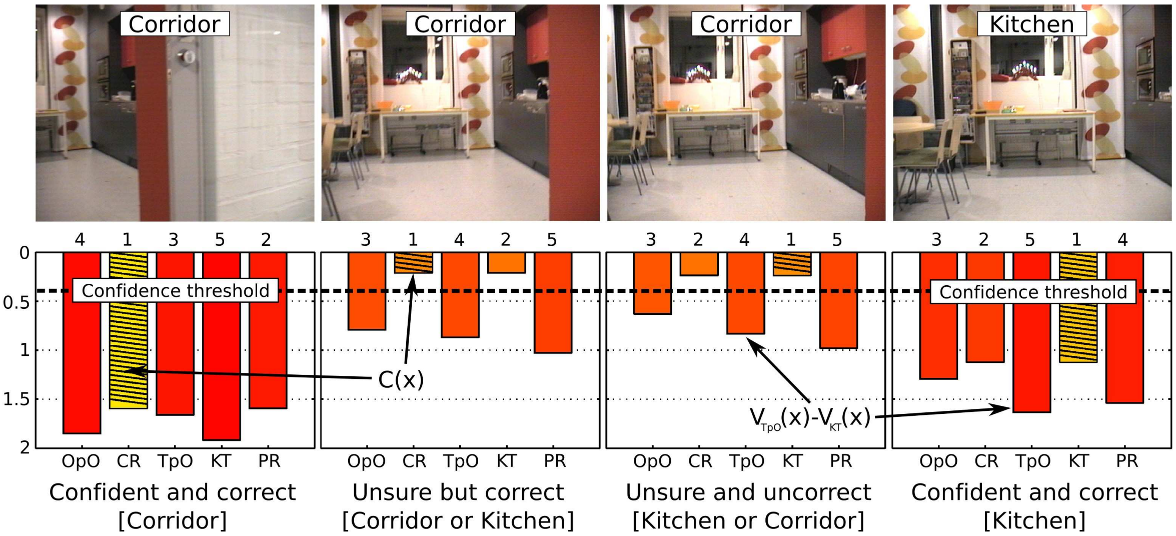
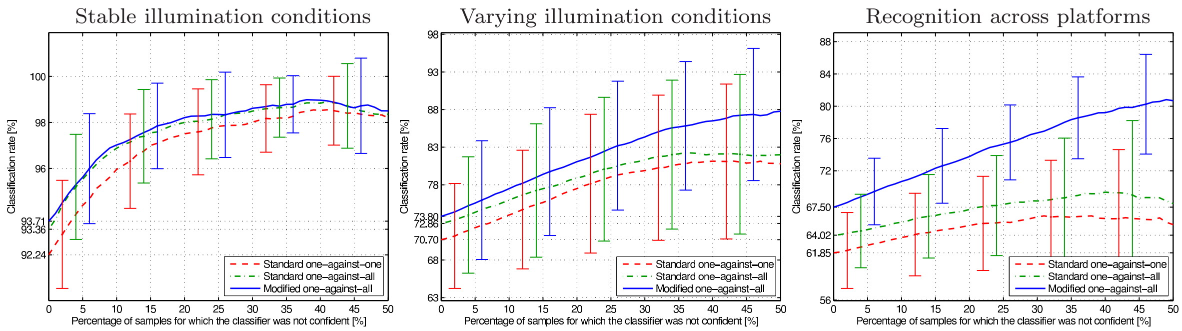
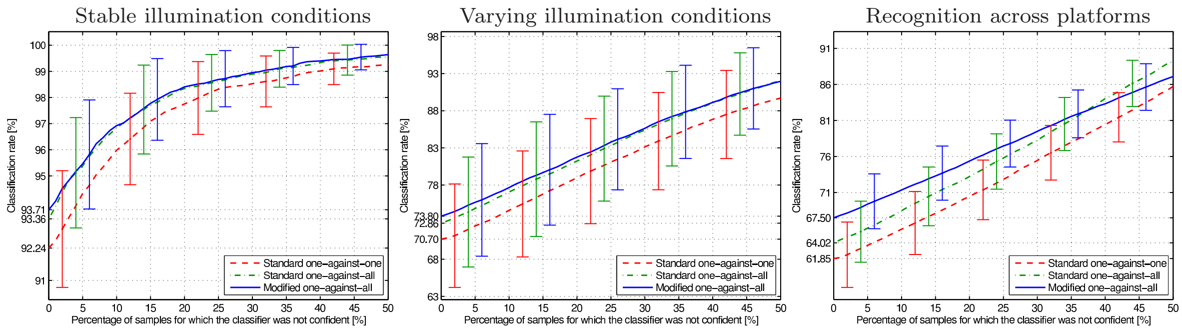


Figure 3: Real confidence estimates obtained using the modified OaA algorithm for four images acquired by the robot Minnie turning from the corridor towards the kitchen. According to the laser-based localization system used as ground truth, the first three images were acquired in the corridor, while the fourth image was already captured in the kitchen. The bar charts show the ranking of hypotheses (top axis), the estimated confidence of the decision (shaded bar), and the difference between the score for the best hypothesis and the others (remaining bars). For the confidence threshold set as shown in the figure, we obtain two soft decisions (suggesting the correct hypothesis) for the cases of lowest confidence.

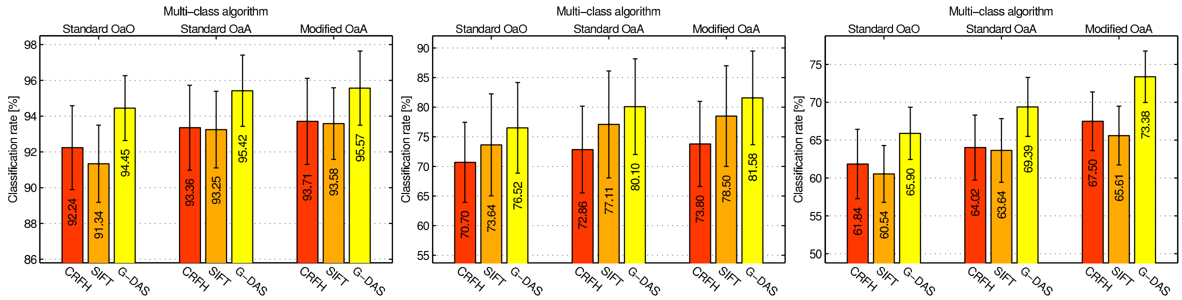


(a) Classification rates for confident samples only



(b) Classification rates including all hypotheses below confidence threshold

Figure 4: Results of evaluation of the three confidence estimation algorithms on three types of problems.



(a) Stable illumination conditions (b) Varying illumination conditions (c) Recognition across platforms

Figure 5: Average results for G-DAS based on different multi-class extensions, for three different series of experiments.

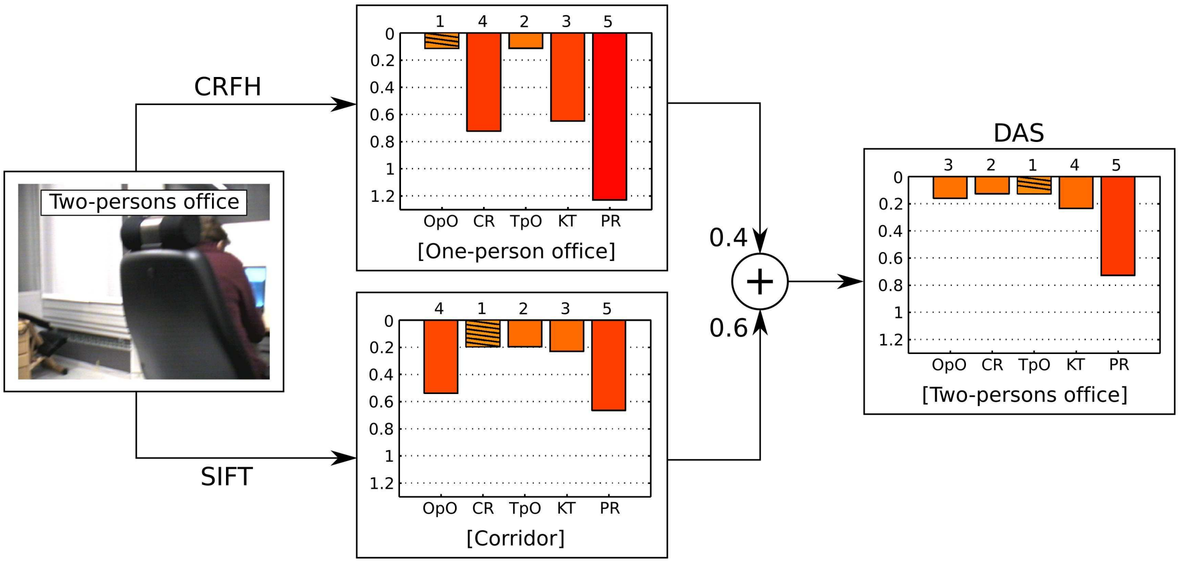
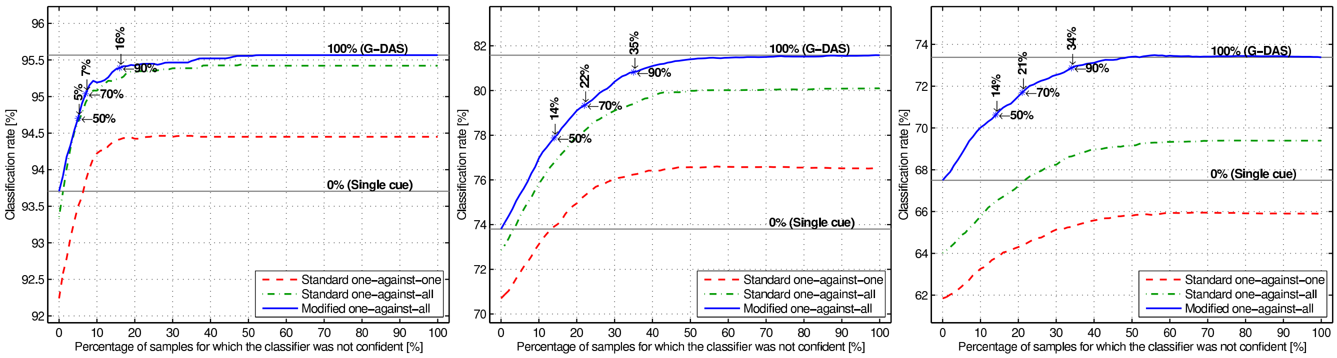


Figure 6: An example of test image misclassified by using a single cue, but classified correctly by using G-DAS with modified OaA multi-class extension (see Fig. 3 for an explanation of the bar charts).



(a) Stable illumination conditions (b) Varying illumination conditions (c) Recognition across platforms

Figure 7: Dependencies between the average classification rates obtained for the confidence-based cue integration strategy with CRFH used as a primary cue and the percentage of test samples for which both cues were used. The horizontal lines indicate the performance of CRFH only and G-DAS.