Int J Comput Vis (2010) 90: 45–61
DOI 10.1007/s11263-010-0336-8

Learning Photometric Invariance for Object Detection

Jose M. Álvarez · Theo Gevers · Antonio M. López

Received: 15 June 2009 / Accepted: 21 March 2010 / Published online: 22 April 2010

Abstract Color is a powerful visual cue in many computer vision applications such as image segmentation and object recognition. However, most of the existing color models depend on the imaging conditions that negatively affect the performance of the task at hand. Often, a reflection model (e.g., Lambertian or dichromatic reflectance) is used to derive color invariant models. However, this approach may be too restricted to model real-world scenes in which different reflectance mechanisms can hold simultaneously.

Therefore, in this paper, we aim to derive color invariance by learning from color models to obtain diversified color invariant ensembles. First, a photometrical orthogonal and non-redundant color model set is computed composed of both color variants and invariants. Then, the proposed method combines these color models to arrive at a diversified color ensemble yielding a proper balance between invariance (repeatability) and discriminative power (distinctiveness). To achieve this, our fusion method uses a multiview approach to minimize the estimation error. In this way, the proposed method is robust to data uncertainty and produces properly diversified color invariant ensembles. Further, the proposed method is extended to deal with temporal data by predicting the evolution of observations over time.

Experiments are conducted on three different image datasets to validate the proposed method. Both the theoretical and experimental results show that the method is robust against severe variations in imaging conditions. The method is not restricted to a certain reflection model or parameter tuning, and outperforms state-of-the-art detection techniques in the field of object, skin and road recognition. Considering sequential data, the proposed method (extended to deal with future observations) outperforms the other methods.

Keywords Object detection · Color models · Learning · Photometric invariance · Combining classifiers · Diversified ensembles

1 Introduction

Color is a powerful visual cue in many computer vision applications such as image segmentation, object recognition and scene classification. Most of the existing color models depend on the imaging conditions under which the image is recorded (such as illumination and camera viewpoint). Varying imaging conditions may disturb the measured color model values and hence the task at hand. Although reflection models (e.g., Lambertian or dichromatic reflectance) are used to derive color invariant models (Finlayson et al. 2006; van de Sande et al. 2008), these reflection models may be too restricted to model real-world scenes in which different reflectance mechanisms can hold simultaneously.

To avoid the requirement of explicit reflection models, a combining strategy is proposed here to obtain photomet-

J.M. Álvarez (✉) · A.M. López
Computer Vision Center and Computer Science Dpt., Universitat
Autònoma de Barcelona, Edifici O, Campus UAB, 08193
Bellaterra (Cerdanyola) Barcelona, Spain
e-mail: jalvarez@cvc.uab.es

A.M. López
e-mail: antonio@cvc.uab.es

T. Gevers
Intelligent Systems Lab. Amsterdam, University of Amsterdam,
Kruislaan 403, 1098 SJ, Amsterdam, The Netherlands
e-mail: th.gevers@uva.nl

T. Gevers
Computer Vision Center, Edifici O, Campus UAB, 08193
Bellaterra (Cerdanyola) Barcelona, Spain

ric invariance. In general, combining multiple classifiers (e.g., color descriptors) that consider the differences between their components is a powerful technique to improve the performance of single classifiers (Brown et al. 2005; Kittler et al. 1998; Kuncheva 2004). In this paper, the measure of disagreement between components is referred as diversity. A promising subset of combining strategies are those using diversity in the process of defining the ensemble (Kuncheva 2004). For instance, Melville and Mooney (2005) consider diversity as the disagreement of an ensemble member with the ensemble's prediction to learn ensembles based on positive and negative data. Jacobs (1995) propose a minimum variance estimator where the estimated aggregate has a variance at most as large as the variance of any of the input features. Stokman and Gevers (2007) uses the Markowitz diversification criterion (Markowitz 1959) in the process of defining the ensemble. The method assumes that each descriptor can be characterized by an unimodal distribution and computes the best combination which provides maximal feature discrimination. However, in practice, the distribution of the training data is often not unimodal leading to estimation errors which are maximized by the quadratic optimization technique used to compute the ensemble (Scherer 2002).

For a given combination strategy, proper selection of its components is important to improve the performance of the strategy. The ideal situation would be a set of classifiers with uncorrelated errors. Then, these classifiers could be combined to minimize the effect of these failures. In fact, the combination of a set of similar classifiers will not outperform the individual members (Kuncheva 2004). The improvement that can be obtained by selecting appropriate classifiers can even be larger when the method uses a learning step to adapt to the specific classification problem (e.g., Boosting, Bagging and Random forests). To facilitate the learning procedure, systems use training data corresponding to the object to be recognized (i.e., positive examples) and for instance background (i.e., negative examples). Systems using only positive data within the training step are more desirable since obtaining a comprehensive representation of negatives or unknown universe is often unfeasible. In addition, if negative data is not chosen properly this may lead to lower classification accuracy (Tax and Duin 2002).

Therefore, in this paper, we aim to derive color invariance by learning from positive examples of color models to obtain diversified color invariant ensembles. The training examples should include a broad range of varying imaging conditions under which the object/image is recorded. An orthogonal and non-redundant color model set is first computed composed of both color variants and invariants. Then, the proposed method combines these color models to arrive at a diversified color ensemble yielding a proper balance between invariance (repeatability) and discriminative power (distinctiveness). To achieve this, the method uses a multi-view approach to minimize the estimation error. In addition, the contribution of each observation is estimated using a Monte Carlo simulation. In this way, the method is more robust to data uncertainty and produces properly diversified color invariant ensembles. Finally, the method is extended to deal with sequential data. Time series modeling is included in the method to predict the evolution of observations over time. To this end, a weighting scheme is used to incorporate the dynamics of observations over time. The ensemble is periodically updated considering the new data available and its order in time. To illustrate the extension of the method, an example experiment using synthetic data is provided. Further, the extension to time series is applied to real-world videos for the purpose to object detection in videos with varying imaging conditions.

The paper is organized as follows. First, in Sect. 2, the multi-view fusion scheme is introduced. In Sect. 3 the method is extended to consider the evolution of data over time. Color-based region detection is outlined in Sect. 4. Then, in Sect. 5, experiments are presented and the results are discussed. Finally, conclusions are drawn in Sect. 6.

2 Multi-view Combination of Observations based on Diversified Ensembles

In this paper, we aim to model a homogeneously colored image region (object) recorded under varying imaging conditions (views) by combining different color models (observations). At each view, the image region contains multiple pixels (samples of an observation). Then, we propose to model the color of the region using a single value (expected color $E[\xi_O]$) with small deviations from this value (σ_O) by:

$$O = E[\xi_O] \pm \sigma_O. \tag{1}$$

Table 1 Definitions and correspondence between symbols and color-related terms

Definitions	
Abstract terms	Color terms
Object O	Homogeneously-colored image region.
View	Image region recorded under a different imaging condition (i.e., illumination, shading).
Object representation ξ_i	Expected value using the i-th color model *independent* of the imaging conditions.
Observation $\tilde{\xi}_{ij}$	Expected value using the i-th color model for the j-th view.
Samples of observation ξ_{ijl}	Pixel values used to estimate the i-th color model data distribution for the j-th view.

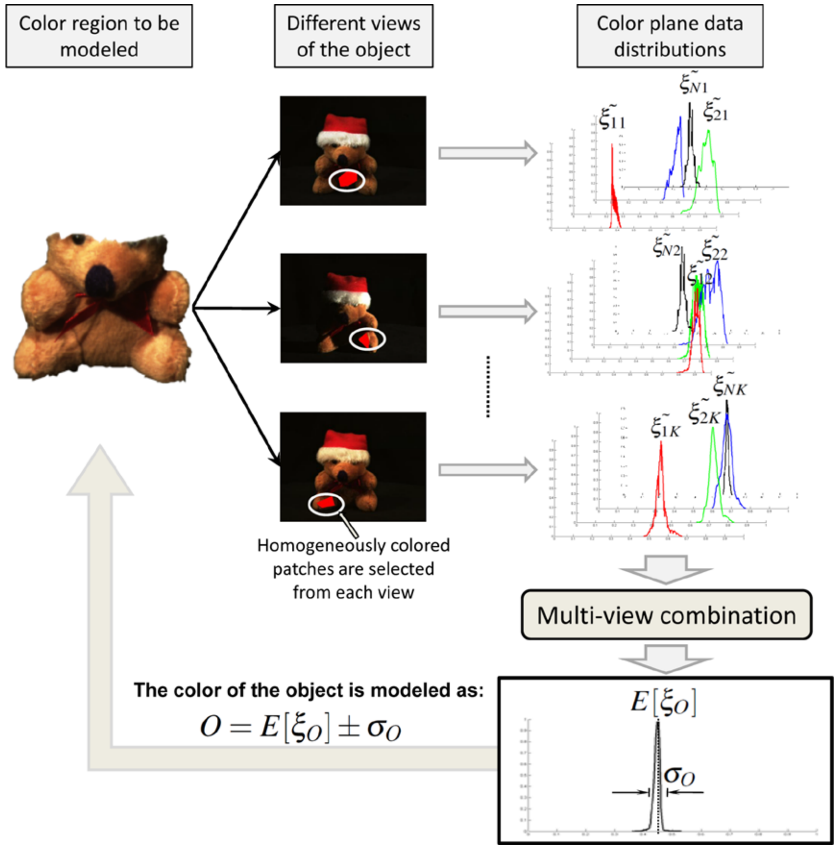

Color region to be modeled
Different views of the object
Color plane data distributions
$\tilde{\xi}_{N1}$
$\tilde{\xi}_{11}$
$\tilde{\xi}_{21}$
$\tilde{\xi}_{N2}$
$\tilde{\xi}_{22}$
$\tilde{\xi}_{12}$
$\tilde{\xi}_{NK}$
$\tilde{\xi}_{2K}$
$\tilde{\xi}_{1K}$
Homogeneously colored patches are selected from each view
Multi-view combination
The color of the object is modeled as:
$O = E[\xi_O] \pm \sigma_O$
$E[\xi_O]$
σ_O

Fig. 1 The color of an image region is modeled combining the information of different color models from different views

Definitions are given in Table 1 and the modeling process is illustrated in Fig. 1.

To build the model, it is assumed that L different samples of N different observations for K views of the object are available ($\xi_{ijl}, i \in [1, 2, \dots, N], j \in [1, 2, \dots, K], l \in [1, 2, \dots, L]$).

These instances correspond, for example, to the same object imaged under varying imaging conditions (e.g., shading, highlights and illumination) generating variations of the observations apart from those due to device-dependent recording noise. Further, multiple samples of each observation are provided to reduce the influence of noise. Then, a set of N orthogonal and non-redundant representations of the object is estimated ($\xi_1, \dots, \xi_N$) and the object is modeled by a weighted linear combination of the representative (expected) values of each observation $E[\xi_i]$:

$$E[\xi_O] = \sum_{i=1}^{N} w_i E[\xi_i], \tag{2}$$

where $\mathbf{w} = [w_1, \dots, w_N]$ is the contribution of each observation to the final combination. Further, the standard deviation of the object is obtained by:

$$\begin{aligned}
\sigma_O^2 &= E[(E[\xi_O] - E[\bar{\xi}_O])^2] \\
&= E\left[\left(\sum_{i=1}^{N} w_i E[\xi_i] - \sum_{i=1}^{N} w_i E[\bar{\xi}_i]\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{N} w_i (E[\xi_i] - E[\bar{\xi}_i])\right)^2\right] \\
&= E\left[\left(\sum_{i,j=1}^{N} w_i w_j (E[\xi_i] - E[\bar{\xi}_i])(E[\xi_j] - E[\bar{\xi}_j])\right)^2\right] \\
&= \sum_{i,j=1}^{N} w_i w_j \sigma_{ij} \\
&= w \Sigma w^T,
\end{aligned} \tag{3}$$

where Σ is the covariance matrix representing the existing relations between observations when the viewing conditions are changing.

To estimate the representative (expected) values of each observation $E[\xi_i]$, a multi-view framework is proposed. This framework characterizes the information available from each observation using two different stages. First, the central value of each observation for each view (ξ_{ij}) is computed using the data distribution of the samples available (ξ_{ijl}). In particular, we use the mode of the samples available for the j-th view of the i-th observation ($\tilde{\xi}_{ij}$). Using this central value, the algorithm minimizes the influence of skewed distributions, thus, minimizes the estimation error. Second, the expected value of an observation given the values of different views $E[\xi_i]$ is estimated assuming that each available view has the same probability of appearing. In particular, the mean value of central values for each view is considered as follows:

$$E[\xi_i] = \frac{1}{K}\sum_{j=1}^{K}\tilde{\xi}_{ij}. \tag{4}$$

What remains is the estimation of w_i. A proper combination of observations leads to a model for which the expected value of the object ($E[\xi_O]$) is close to a reference value ($E[\xi_{O_R}]$) and for which its variance is minimized. In this way, the combination reduces the deviations from the expected value due to varying viewing conditions. This reference value is, for example, the value which is obtained when ideal acquisition conditions are obtained. Hence, computing w_i can be posed as an optimization problem formulated as follows,

$$\text{minimize} \quad \sum_{i,j=1}^{N} w_i w_j \sigma_{ij} \tag{5}$$

subject to $E[\xi_O] \geq E[\xi_{O_R}]$,

$$\sum_{i=1}^{N} w_i = 1, \tag{6}$$

where the full combination constraint has been added. That is, the contributions of each observation must sum up to one.

Quadratic optimization techniques (Boyd and Vandenberghe 2004) can be applied to solve (5) and provide a set of optimum solutions (efficient ensembles) called the efficient frontier (Scherer 2002). That is, the efficient frontier contains different values of $E[\xi_O]$ and associated weights which minimize the corresponding σ_O. However, quadratic optimization techniques tend to select components with attractive characteristics so components with the less appealing features are not selected. These are the cases in which the estimation error is likely to be maximal (Michaud and Michaud 2008; Scherer 2002). Therefore, in order to deal with the estimation error and improve the diversity of the ensemble, in this paper, a resampling technique is proposed. This resampling technique uses a Monte Carlo simulation to obtain a set of efficient ensembles called resampled frontier (Michaud 1998). Ensembles lying on this resampled frontier are composed of weight vectors obtained as the average of the efficient frontiers given a certain expected value. The performance of resampled efficient ensembles is better than the performance of those ensembles obtained using quadratic optimization techniques (Michaud and Michaud 2008; Usmen 2003).

Finally, the most appropriate ensemble is selected from the set of ensembles lying on the efficient frontier using the Sharpe Ratio (*SR*) (Dowd 1998). This ratio is a single statistical performance measure of variance-adjusted expected return defined as follows:

$$SR = \frac{E[\xi_O]}{\sigma_O}. \tag{7}$$

The highest *SR* corresponds to the ensemble in the frontier obtaining best performance. If a benchmark ensemble exists ξ_R, the performance of an ensemble in the frontier (P_e) is computed as follows:

$$P_e = \frac{1}{(E[\xi_O] - \xi_R)\sigma_O}, \tag{8}$$

The highest performance corresponds to the most appropriated ensemble.

The above computation of weights and the ensemble selection method are summarized as follows:

1. Estimate the efficient frontier using the training data and quadratic programming techniques. This frontier is composed of ensembles varying from minimum-variance to the maximum expected value ensembles. Divide the difference between the minimum and maximum return in m ranks.
2. Estimate the variance-covariance matrix, Σ, and expected values, $E[\xi_i]$, of the training data,
$$E[\xi_i] = \frac{1}{K}\sum_{j=1}^{K}\tilde{\xi}_{ij}, \tag{9}$$
$$\Sigma = (\sigma_{i,j}), \tag{10}$$
where K is the number of views.
3. Resample, using the training inputs in Step 2, taking D draws for the input multi-variate distribution. The number of draws D reflects the degree of uncertainty in the training data. Compute a new variance-covariance matrix from the sampled series. Estimation error will result in different variance-covariance matrices and mean vector from those in Step 2.
4. Compute the efficient frontier for the inputs derived in Step 3. Calculate the optimal ensemble weights for m equally distributed points along the frontier.
5. Repeat Step 3 and Step 4 P times. Calculate the averaged ensemble weights for each observation,
$$\overline{w}_i^{resampled} = \frac{1}{P}\sum_{im=1}^{P} w_{im}, \tag{11}$$
where w_{im} denotes the weight vector for the m-th ensemble along the frontier for the i-th observation.

6. Evaluate the frontier of averaged ensembles by the variance-covariance matrix from the original training data to obtain the resampled frontier.
7. Select the ensemble from the frontier which exhibits the highest performance (see (7) or (8) as required).

3 Temporal Ensemble Adaptation

An important characteristic of any learning system is its adaptation to newly obtained data. The previous section provides an optimal weighted combination of observations based on training samples representing different views of the same object (Table 1). However, no information regarding the order of these observations was considered to build the model. In this section, the proposed model is extended to take into account the evolution of observations over time (e.g., from still images to videos). The key idea is to include temporal information in the estimation of the parameters $E[\xi_1], \dots, E[\xi_N]$ and Σ. These parameters are computed considering each view of a given observation provides the same information to the final ensemble. However, due to the dynamic nature of data sequences, current observations should be taken into account more prominently than distant ones. In this way, the modification of the algorithm consists of using time series analysis to predict the expected values of observations rather than considering simple averages over views.

To express the dynamic structure of the data (observations and ensembles), a weighting process is used. Further, the dynamic structure of the variance within observations is also considered. There are two models to deal with these kind of variations: exponentially weighted moving average (EWMA) and generalized autoregressive conditionally heteroscedastic (GARCH). Both models assume that serial correlation is present in the dynamics of the observations. As a result, both models assign higher weights to recent values than the older ones. In particular, in this paper, EWMA model is used mainly due to its simplicity (less parameters to estimate) and the ability to cope with changes in standard deviation of the incoming data (Best 1998; Tse 1991).

EWMA uses a decay factor that weighs the change of each past observation. More recent observations receive higher weights than older ones. Using EWMA, the input parameters of the optimization process are derived as follows:

$$E[\xi_i] = \frac{1}{\sum_{j=1}^{K} \lambda^{j-1}} \sum_{j=1}^{K} \lambda^{j-1} \tilde{\xi}_{ij}, \tag{12}$$

$$\Sigma = (\sigma_{nm}) = (1-\lambda) \sum_{j=1}^{K} \lambda^{j-1} (\tilde{\xi}_{nj} - E[\xi_n])(\tilde{\xi}_{mj} - E[\xi_m]), \tag{13}$$

where λ is the decay factor. This factor determines both the degree of weighting of recent observations and also the speed with which the volatility measure will return to a lower level after a large return. A lower decay gives a higher weighting to recent values. K is the number of past observations unlike the previous section where K was the number of different views available for each observation. Parameter K can be set to infinity since the weighting procedure will rapidly reduce to zero for distant observations. Since $0 < \lambda < 1$, $\lambda^n \to 0$ when $n \to \infty$, the model will eventually place a zero weight on observations far in the past.

3.1 Synthetic Data Example

To illustrate the effectiveness of the time-adaptation process, the following experiment using synthetic data is conducted. The aim of the experiment is to evaluate the theoretical improvement achieved when the evolution of observations is considered to construct the ensemble. The data set consists of five different sequences of observations (Fig. 2). Each sample in the dataset corresponds to the central value of observations at a certain instant of time ($\tilde{\xi}_{ij}$) where $i \in [1, \dots, 5]$ and $j \in [1, 2, \dots, K]$. The dataset contains a total of 2200 samples, 440 for each observation ($K = 440$). The first 40 samples from each observation are used as training data to build an ensemble without considering the temporal information.

Different updating techniques are evaluated. First is the so-called sample and hold method where an ensemble is built using training data which will not be updated when new information is available. Further, different decay factors are considered, ranging from $[0, 0.1, \dots, 1]$, where $\lambda = 1$ corresponds to equally weighting all the historical samples. For this, a look-back period S has to be defined. For other values of λ, there is no need to fix this.

The evaluation consists of assessing the adaptation of the ensemble to the dynamics of the data. Hence, the results achieved by the ensemble process are compared against the ideal result or benchmark for each process. Note that in the case of simulated data, the true values of the future ensemble are known, and thus, the optimal (as opposed to estimated) mean-variance ensemble can be computed. This ensemble is used as benchmark for each combination strategy.

The tracking error (ψ) (Rasmussen 2003) is used as a performance measure. This measure denotes how closely an ensemble follows the value to which it is benchmarked. Hence, the tracking error of an ensemble E relative to some benchmark B over K time periods is:

$$\psi = \frac{1}{(K-1)} \sum_{t=1}^{K} (D_t - \bar{D})^2, \tag{14}$$

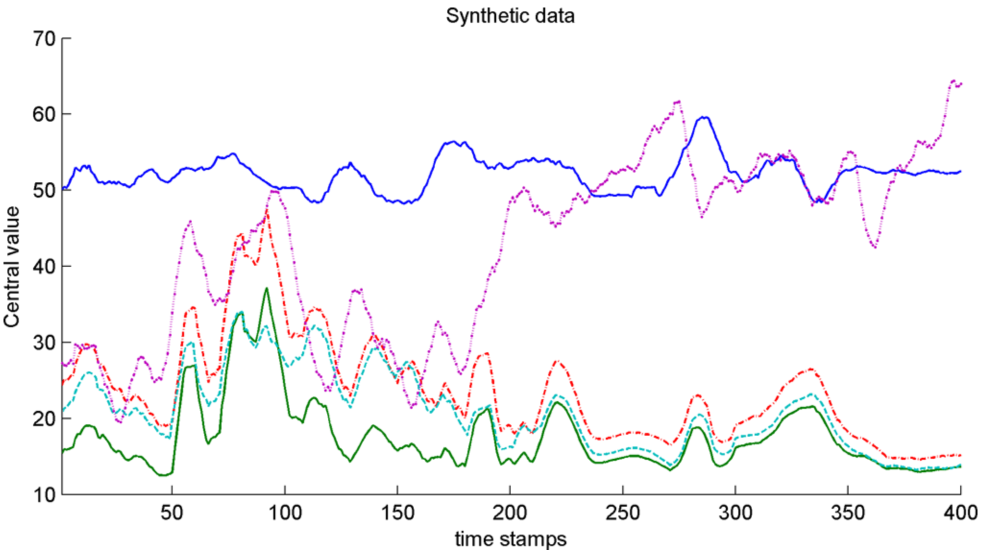


Fig. 2 Synthetic data used to validate the capabilities of the tracking algorithm

where $D_t = (E[\xi_{Ot}] - E[\xi_{Bt}])$ is the difference between the ensemble value and the incoming value of the benchmark and $\bar{D} = \frac{1}{K}\sum_{t=1}^{K} D_t$ is the mean of these differences. The lower ψ is, the closer the ensemble is to its benchmark. Further, to test the statistical significance of the results, the historical *SR* (S_h) or Ex-post *SR* (Sharpe 1994) is used. This ratio is derived from the *SR* which is a measure of expected value per unit of variance in an ensemble. *SR* is used to characterize how well the expected value of an ensemble compensates for the existing variance. Further, S_h measures the performance of the ensembles against a benchmark over a period of time,

$$S_h = \frac{\bar{D}}{\sigma_D}, \tag{15}$$

where σ_D is the square root of ψ. This measure is closely related to the t-statistic. The t-statistic is equal to the *SR* multiplied by the square root of K (Sharpe 1994). However, this measure requires proper interpretation since the original measure aims at maximizing the numerator while minimizing the denominator. Thus, to properly use this measure, the difference between the expected value and the current value is reversed to express the same behavior. Hence, the highest S_h is the highest historical performance. A summary of the results is listed in Table 2. As shown, a higher performance (lower tracking error) is achieved when the composition of the ensemble is updated over time. The improvement is higher when the exponentially weighted scheme is adopted. However, as expected, the performance depends on λ: the relevance of current samples with respect to distant ones. In any case, updating the ensemble improves the performance of the algorithm except when all historical data is equally considered. This is an expected performance as the expected value is predicted considering information which is rarely related to the incoming one.

Table 2 Tracking error (ψ) and historical Sharpe ratio (S_h) comparison for different values of time adaptation configurations. The higher S_h the better performance and the lower ψ the better performance

EWMA	ψ	S_h
Sample and hold	6.2068	2.8435
$\lambda = 1, S = 30$	10.7631	11.2124
$\lambda = 1, S = 50$	0.8631	10.0124
$\lambda = 1, S = \infty$	3.3215	4.6842
$\lambda = 0.9$	0.4892	13.5402
$\lambda = 0.8$	0.2205	20.5731
$\lambda = 0.7$	0.2497	19.4215
$\lambda = 0.6$	0.2546	19.2591
$\lambda = 0.5$	**0.1551**	**24.7231**
$\lambda = 0.4$	0.2067	21.3939
$\lambda = 0.3$	0.3748	15.8704
$\lambda = 0.2$	0.4171	15.0195
$\lambda = 0.1$	0.2476	19.4857

4 Application to Color-based Region Detection

In this section, the method presented in the previous section is applied to color-based region detection, that is, the detection of object patches in images recorded under varying imaging conditions using a set of color models composed of both color variant and invariant models. The goal is to derive color invariance by learning from color models to obtain diversified color invariant ensembles.

To this end, every possible transformed color model is considered as an observation of the same object (color-region) and each view corresponds to a different imaging conditions such as lighting, viewing and illumination variations. Further, each pixel within the region corresponds to different sampling values of the observation. Hence, the proper interpretation of the algorithm is as follows: O is the data distribution of the final combination of color (invariant) planes/models and $E[\xi_O]$ and σ_O its central value and variance respectively. $E[\xi_i]$ is the expected value of the i-th color (invariant) plane estimated using the multi-view procedure. That is, considering first the data distribution from pixels of each view and then the average value of the views. Finally, w_i denotes the contribution of the i-th color model to the final ensemble.

During *training* (i.e., estimating $E[\xi_O]$, σ_O and w_i), the following steps are performed:

- Select a set of training images containing the object to be detected imaged under different acquisition conditions (e.g., varying illumination).
- Select a region of interest for each training image (i-th) and for each color model (j-th) estimate $\bar{\xi}_{ij}$ using the data distribution of pixels in the training region.
- Estimate the correlation matrix Σ of these values. This matrix contains information regarding the relative variations of each color model when the acquisition conditions vary.
- Estimate the weights $\mathbf{w}$ using the Monte Carlo method considering the central value of each color model for each view and the covariance matrix as input data.
- Compute $E[\xi_O]$ and σ_O using (2) and (3) respectively.
- Finally, compute the SNR_O ratio of the model as follows:

$$SNR_O = \frac{E[\xi_O]}{\sigma_O}. \qquad (16)$$

Then, during *classification*, the following steps are performed:

- Convert the image into the color models (the same as during training) and apply the weights $\mathbf{w}$ obtained in the training phase to combine them. This leads to a grey-level image.
- Estimate the signal to noise ratio, *SNR*, by dividing, at each pixel, the local mean value by the local standard deviation. The *SNR* is estimated using a rectangular region ($M \times N$ pixels) at each pixel.
- Compute the error between the SNR_O and the local *SNR* for each pixel. The lower the error, the more similar the colors are.
- Threshold the error image e to obtain the final binary mask C:

$$C(x,y) = \begin{cases} 1 & \text{if } e(x,y) < T, \\ 0 & \text{otherwise.} \end{cases} \qquad (17)$$

The appropriate value of T is obtained using automatic thresholding techniques such as the isodata method (Ridler and Calvard 1978).

Finally, if *temporal adaptation* is required, the following steps are performed:

- Use the proposed classification procedure to classify pixels in the first image.
- Use the current result to estimate the central value of each color model for that frame. Add these central values to the historical data.
- Estimate input parameters to the optimization process (Σ and $E[\xi_1], \dots, E[\xi_N]$) using the *EWMA* process outlined in Sect. 3.
- Select the optimal ensemble from the frontier considering the same *SR* and reference as in the initial training stage.
- Use the new ensemble to process the incoming image.

To provide robustness against confounding imaging conditions (e.g., illumination, shading, highlights, and interreflections), different color models exhibiting different photometric invariance properties have been proposed derived from *RGB* color model in Table 3 (van de Sande et al. 2008). For instance, for the dichromatic reflection model, normalized color *rgb* is to a large extent invariant to a change in camera viewpoint, object pose, and the direction and intensity of the incident light. See Table 4 for an overview of color models and their invariance properties. In addition to the models described by van de Sande et al. (2008), the illumination invariant ($\Im$) proposed in Finlayson et al. (2006) is included. This color invariant requires a calibration parameter, the invariant direction which is an intrinsic parameter of the camera. Currently, this invariant direction can be found either by following the calibration procedure outlined in (Finlayson et al. 2006) or, by using a procedure which determines the invariant direction from a single image (Finlayson et al. 2004) or from a set of images (Álvarez et al. 2008). The former consists in acquiring images of a Macbeth color checker under different day time illuminations and then obtaining the invariant direction by analyzing the *log*-chromaticity plot generated from these images. The latter considers the entropy of a single image to compute the invariant direction. Then, the method consists in generating invariant-images using all the possible invariant directions within a range. The optimum direction is the one minimizing the entropy of its corresponding illumination-invariant image (Finlayson et al. 2004).

Considering all the color models in Table 4, a set is obtained of both color variants and invariants to achieve both distinctiveness and repeatability respectively. The next step is to obtain a non-redundant subset. Covariance matrix Σ provides information about correlation between color models. This analysis can be done using principal component analysis (PCA) (Jolliffe 2002). Then, correlation between color models is represented by the loadings of each color model, see Fig. 3. The input data to PCA is the matrix containing the expected values for each view of each color models ($\bar{\xi}_{ij}$). The closer two points are in the loading space, the more correlated they are (and their corresponding color models). The number of principal components depends on

Table 3 Derivation of opponent color space, normalized *rgb*, *HSV* and *CIELab* color spaces from *RGB* values

Opponent Color Space	Normalized *rgb*
$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$r = \frac{R}{R+G+B}$ $g = \frac{G}{R+G+B}$ $b = \frac{B}{R+G+B}$

HSV	CIE *Lab*
$\begin{pmatrix} V \\ V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$ $H = \arctan \frac{V_2}{V_1} \quad S = \sqrt{V_1^2 + V_2^2}$	$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.490 & 0.310 & 0.200 \\ 0.177 & 0.812 & 0.011 \\ 0.000 & 0.010 & 0.990 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$ $L = 116(\frac{Y}{Y_0})^{1/3} - 16$ $a = 500\left[(\frac{X}{X_0})^{1/3} - (\frac{Y}{Y_0})^{1/3}\right]$ $b = 200\left[(\frac{Y}{Y_0})^{1/3} - (\frac{Z}{Z_0})^{1/3}\right]$ X_0, Y_0 and Z_0 are the coordinates of a reference white point.

Table 4 Invariance of color models (derived in Table 3) for different types of lighting variations i.e., light intensity (LI) or light color (LC) change and/or shift (van de Sande et al. 2008). Invariance is indicated with '+' and lack of invariance with '−'

Taxonomy of color spaces	LI change	LI shift	LI change & shift	LC change	LC change & shift
RGB	−	−	−	−	−
O_1,O_2	−	+	−	−	−
O_3, intensity, L	−	−	−	−	−
Saturation (S)	−	+	+	−	−
Hue (H)	+	+	+	−	−
r, g, a, b	+	−	−	−	−
$\Im$	+	+	+	+	+

the data and the amount of variation. The selection of color models which represent each cluster (e.g., S or b in Fig. 3) is computed by the Hartigan's test for unimodality (Hartigan and Hartigan 1985). In this way, an orthogonal (variant/invariant) and non-redundant (decorrelated) color model subset is obtained which will be used as input of the proposed method as explained and tested in the next section.

5 Experiments

In this section, the proposed algorithm is applied to three different databases: (1) the Amsterdam Library of Object Images (ALOI) (Geusebroek et al. 2005), (2) Caltech Face database (Weber 1999) and (3) a road sequence taken by an on-board camera. The goal of the first experiment on the ALOI dataset is to detect object regions under varying imaging conditions. The second experiment consists of detecting skin to detect faces in the Caltech image dataset. The aim of the last experiment is to detect roads under uncontrolled imaging conditions. These experiments are conducted using thirteen color models ($\Im$, R, G, B, r, g, O_1, O_2, L, a, b, S, V). The third opponent color O_3 is excluded since it provides intensity information which is already provided by V. Further, the hue component H from the *HSV* color space is excluded due to its instability close to the achromatic axis (Kender 2005). The calibration required to compute $\Im$ is done using the approach proposed in (Álvarez et al. 2008). Finally, the reference white point for deriving *CIELab* color space is set to the D65 white point ($X_0 = 0.9505$, $Y_0 = 1.0000$, $Z_0 = 1.0888$) (Wyszecki and Stiles 1982).

5.1 Error Measures

Quantitative evaluations are provided using pixel-based measures, see Table 5, from which the following error measures are computed: quality, detection accuracy, detection rate and effectiveness, see Table 6. Each of these measures

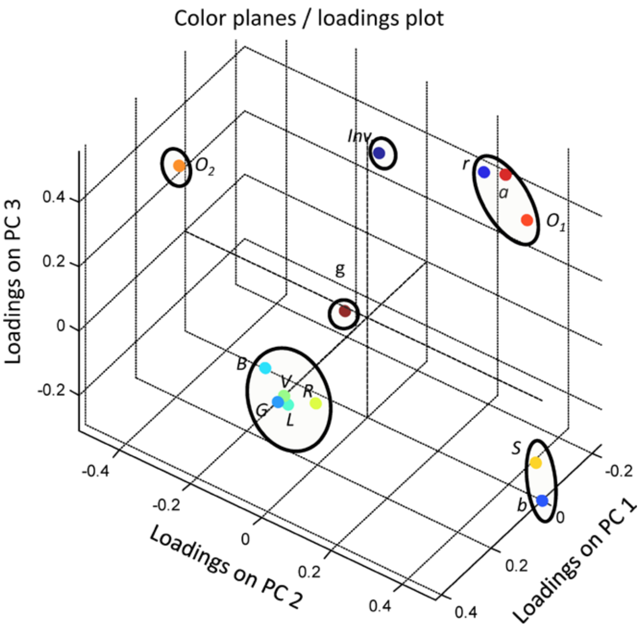


Fig. 3 PCA is used to reduce redundancy within the training data. The analysis is done using the loadings plot of each color model. This example corresponds to the training set from the face database

Table 5 The contingency table. Algorithms are evaluated based on the number of pixels correctly and incorrectly classified

Contingency table	Ground truth		
		Non-target	Target
Detection	Non-target	TN	FN
Result	Target	FP	TP

Table 6 Pixel-wise measures used to evaluate the performance of the different algorithms. These measures are defined using the entries of the contingency table (Table 5)

Pixel-wise measure	Definition
Quality ($\hat{g}$)	$\hat{g} = \frac{TP}{TP+FP+FN}$
Detection Accuracy (*DA*)	$DA = \frac{TP}{TP+FP}$
Detection Rate (*DR*)	$DR = \frac{TP}{TP+FN}$
Effectiveness (*F*)	$F = \frac{2DADR}{DA+DR}$

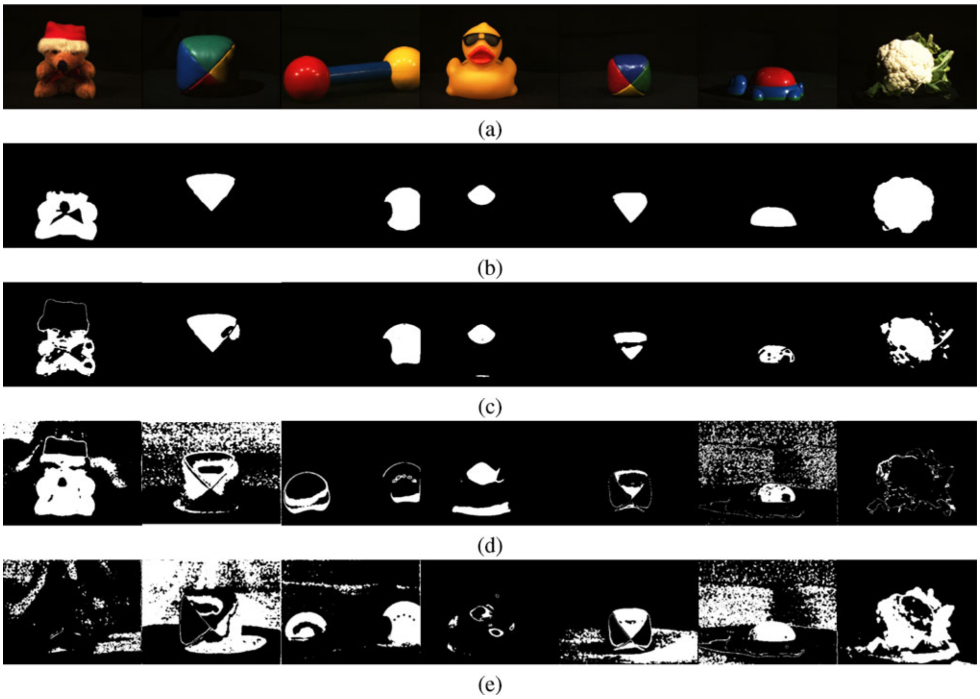
provides a different insight in the performance of a method. Quality takes into account the completeness of the extracted data as well as its correctness. Detection accuracy, also known as precision, is the probability that the result is valid. Detection rate, or recall, is the probability that the ground-truth data is detected. Effectiveness is a single measure that trades-off the detection accuracy versus detection rate. Further, the performance of our method is compared, on each dataset, to existing algorithms. Pair-wise comparisons between algorithms are computed by the Wilcoxon statistical significance test (Wilcoxon 1945).

5.2 Object Region Detection in Controlled Environments: Still Images

Objects are taken from the Amsterdam Library of Object Images (ALOI) (Geusebroek et al. 2005). The objects are captured under different viewing angles, illumination angles, and illumination color. The lights were chosen to be representative of the spread of common illuminants. The goal of this experiment is to detect object regions. The experiment is conducted on each image (36 per object) of seven different objects (1, 25, 55, 62, 85, 142, 877). These objects are selected for the ease of illustration (Fig. 4a) according to two different criteria: objects having different clearly defined color regions and diversity on the object region colors. The corresponding ground-truth is generated by manually segmenting all these images (Fig. 4b).

Fifteen regions from 15 images of each object are randomly selected to train the algorithm representing less than 5% of the total amount of pixels of each region. Although these regions comprise the same object under different lighting conditions, each region contains pixels from an homogeneously colored patch. The data distribution of each patch is approximately unimodal. The set of photometric orthogonal and non-redundant color models is computed using the (PCA) procedure described in Sect. 4. As a result, weights are obtained and listed in Table 7. Some color models exhibiting high weights (i.e., $\Im$, V and L) do not contain chromatic information. Hence, the color of the object's region is reflected in other color models with high weight such as R, B, g or O_2. Example detection results for each object are shown in Fig. 4c.

For comparison, two different weighting algorithms have been implemented: the minimum variance (Jacobs 1995) and the single-view fusion scheme (Stokman and Gevers 2007). The same pixels are used to train all algorithms and all the images of each object have been used for testing. A number of representative results are shown in Fig. 4d and Fig. 4e. A summary of quantitative results is reported in Table 8 for a single object providing insight into the method and in Table 9 for overall average performance evaluation. Analyzing a single object, the proposed multi-view approach outperforms the other approaches in terms of overall performance (quality, detection accuracy and F measure). Nevertheless, it does not outperform the single-view approach in terms of detection rate. Hence, the single-view method detects more target pixels at the expense of miss-detected background pixels. Analyzing the average performance for all the objects, the proposed multi-view approach clearly outperforms the other approaches in terms of overall average performance. Thus, it can be concluded that considering only the variance of the training data (i.e., the minimum variance method) is not sufficient to provide a proper model. In fact, single-view methods not only aim at minimizing the variance but they also yield a certain mean value


(a)

(b)

(c)

(d)

(e)

Fig. 4 Example results for different objects from the ALOI database. (**a**) Original images. (**b**) Ground-truth generated manually. (**c**) Results by the proposed method. (**d**) Results obtained using the minimum variance fusion method (Jacobs 1995). (**e**) Results obtained using the single-view fusion method (Stokman and Gevers 2007)

Table 7 Set of weights obtained for the experiments. '–' corresponds to an unselected color model by the PCA procedure

	Teddy-bear	Green-ball	Toy	Beck	Blue-ball	Turtle	Cabbage	Skin	Road
$\Im$	0.595	–	0.304	0.396	0.020	0.017	0.003	−0.017	0.929
R	0.048	0.049	0.020	0.353	–	0.042	0.145	–	–
G	–	0.037	–	–	–	0.213	–	–	–
B	0.328	–	–	–	–	–	0.004	–	–
r	–	0.196	0.275	–	–	0.301	0.005	–	0.157
g	0.276	0.181	0.401	0.439	0.004	0.290	–	0.022	0.342
O_1	–	–	–	–	–	–	–	–	0.266
O_2	–	0.212	–	0.217	0.474	0.101	0.405	0.013	−0.024
L	–	0.041	–	–	–	–	0.017	0.176	–
a	–	0.131	–	–	–	0.035	–	0.652	−0.356
b	–	0.149	–	–	0.489	–	–	0.154	−0.082
S	−0.032	–	–	–	0.013	–	0.421	–	−0.452
V	−0.215	–	–	−0.406	–	–	–	–	0.220

Table 8 Performance of different detection algorithms for the first object from the ALOI database. Bold values indicate maximum performance

	$\hat{g}$	Detection accuracy	Detection rate	F
Minimum variance (Jacobs 1995)	0.156 ± 0.09	0.269 ± 0.12	0.305 ± 0.21	0.259 ± 0.15
Single-view Fusion (Stokman and Gevers 2007)	0.478 ± 0.11	0.627 ± 0.13	$\mathbf{0.789 \pm 0.06}$	0.694 ± 0.14
Multi-view[a]	0.294 ± 0.15	0.419 ± 0.28	0.784 ± 0.24	0.627 ± 0.25
Multi-view (our method)	$\mathbf{0.639 \pm 0.13}$	$\mathbf{0.909 \pm 0.03}$	0.687 ± 0.15	$\mathbf{0.778 \pm 0.11}$

[a] Without color model selection

Table 9 Average performance of different detection algorithms for seven objects from the ALOI database (objects 1, 25, 55, 62, 85, 142, 877). Bold values indicate maximum performance

	$\hat{g}$	Detection accuracy	Detection rate	F
Minimum variance (Jacobs 1995)	0.126 ± 0.14	0.317 ± 0.31	0.212 ± 0.22	0.199 ± 0.19
Single-view Fusion (Stokman and Gevers 2007)	0.286 ± 0.26	0.365 ± 0.32	0.656 ± 0.33	0.389 ± 0.28
Multi-view[a]	0.310 ± 0.23	0.383 ± 0.28	0.599 ± 0.21	0.414 ± 0.26
Multi-view (our method)	$\mathbf{0.514 \pm 0.21}$	$\mathbf{0.750 \pm 0.29}$	$\mathbf{0.703 \pm 0.21}$	$\mathbf{0.655 \pm 0.18}$

[a] Without color model selection

of the ensemble. Hence, the extra information provided by the multi-view approach is important to achieve a proper ensemble. Finally, since training pixels are obtained under different imaging conditions, the behavior of the different color models cannot be captured properly. In contrast, the proposed method is able to model this phenomenon due to the relative variations around the central value in each view and hence outperforming the other methods.

5.3 Skin Detection: Still Images

The second experiment consists of detecting skin pixels of faces from The Frontal Face Image Database of Caltech. This image dataset contains 450 face images taken from 27 different persons under different lighting, expressions and backgrounds. The appearance of the face in these images is clearly influenced due to different illumination, shading, skin tone and so on (Fig. 5). Ground-truth is generated by segmenting manually all the images in the database. The training set is obtained by manually selecting 100 different patches from 100 different (randomly chosen) images. The unimodality test is used to discard unappropriate patches. Finally, 58 patches are used for training representing 1% of the total of facial pixels in the database. Note that, in this experiment, the covariance matrix (Σ) encapsulates variations not only in the illumination conditions but also in the appearance of the object (i.e., skin tone variations) since different instances of the same object-class are considered at the same time. The color model set is computed using the procedure described in Sect. 4. The set of weights obtained are listed in Table 7. These weights reveal a dominance in a and b reflecting pale reddish (i.e., skin). Example results are shown in Fig. 6. For each original image (Fig. 6a) the weighted combination (Fig. 6b) and the skin data distribution (Fig. 6c) are provided. Further, all the skin pixels in the database are collected and the distribution of values of different color models is shown in Fig. 7. For comparison, only one color model from each group in Table 4 is considered. As shown, the proposed method leads to an unimodal distribution of pixels despite light color and skin tone variations. That is, lighting variations are compensated when color models are properly combined. Note that pixel values for other color models are not normally distributed leading to erroneous mean and standard deviation values.

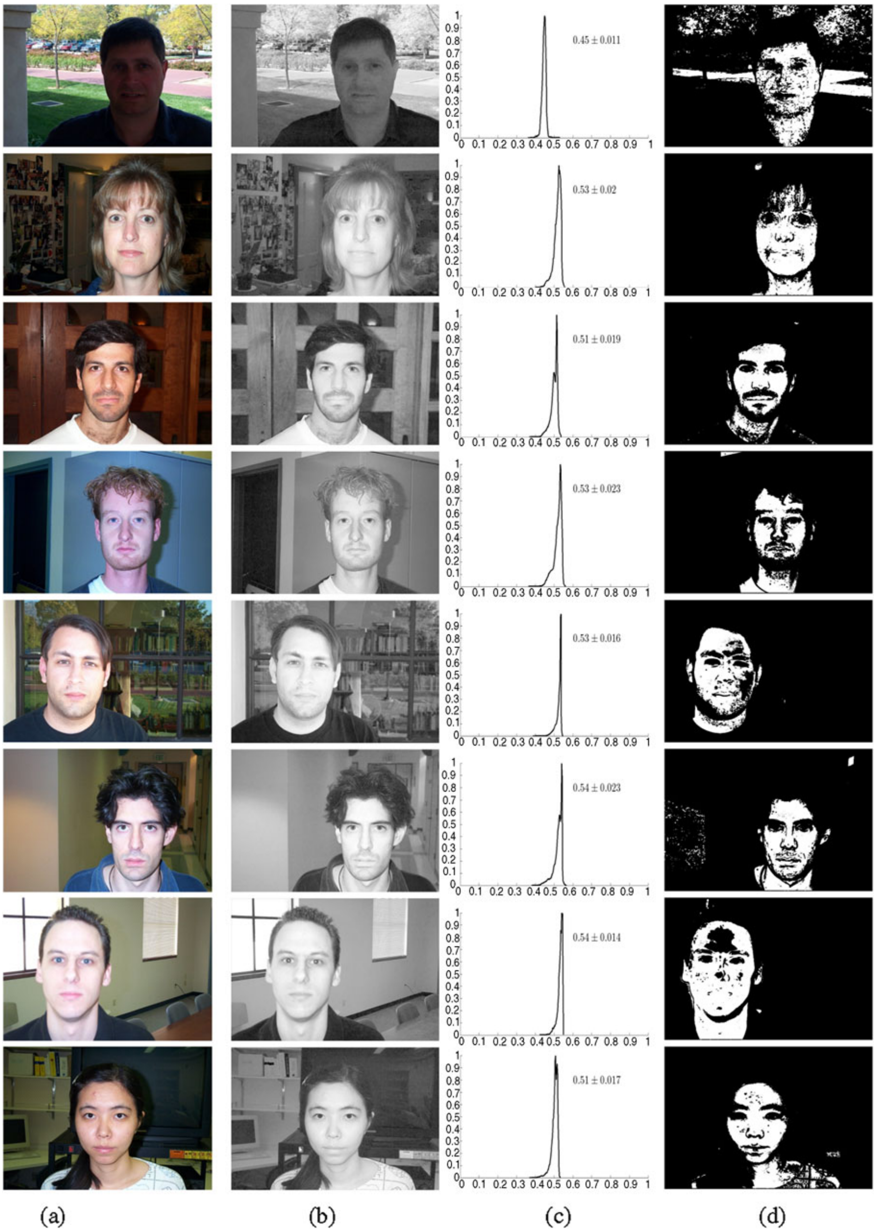
The performance of the method is compared to six different existing skin detection algorithms. Three of them use fixed boundaries in *RGB* (Fleck et al. 1996), *CbCr* (Chai and Ngan 1999) and *HS* (Sobottka and Pitas 1998) color spaces. The fourth is a statistical approach using a mixture of Gaussians in *RGB* space. Note that these methods are particularly designed and fine-tuned to detect skin. The other two methods correspond to the (more generic) fusion schemes proposed by (Jacobs 1995) and (Stokman and Gevers 2007). The same training set is used to train the different detection schemes. A summary of the results is listed in Table 10. Further, the results of the Wilcoxon test are shown in Table 11. The following conclusions can be derived from these results. First, the proposed algorithm outperforms the others in terms of overall performance (quality and effectiveness) except for the *RGB* based method. Nevertheless, the *RGB* based, *HS* and *RGB* statistical method provides better detection rate. This means that these methods provide higher invariance to skin-class variability at the expense of having low discriminative power. Our method outperforms all the others, including *RGB* based method, in terms of detection accuracy. That is, the ratio, between skin pixels which are correctly classified and the number of skin pixels retrieved provided by our method, is higher. This is due to the resulting distribution of skin pixel values (Fig. 7). However, the overall performance of our method is lower than the *RGB* method because of the high variability in both skin appearance and lighting variations. This yields a data distribution in each view which is not unimodal except for very small patches of skin. Further, unobserved lighting conditions and user appearance (during training) shift the skin distribution (Fig. 6c) reducing the performance. Furthermore, although the *RGB*-based method fails in the presence of low intensity (due to illumination and shadows) there are only a few instances of this type i.e., only 3% of images in the image dataset showing severe intensity and shadow changes.

5.4 Road Detection in Video Sequences

Previous experiments considered still images. That is, no temporal information is available. Hence, the optimal en-

Fig. 5 Example images from The Frontal Face Image Database of Caltech (Weber 1999)

Fig. 6 Generic skin detection results (second skin experiment). (**a**) Original image; (**b**) weighted combination of color models; (**c**) distribution of skin pixel values in the image; (**d**) skin detection results



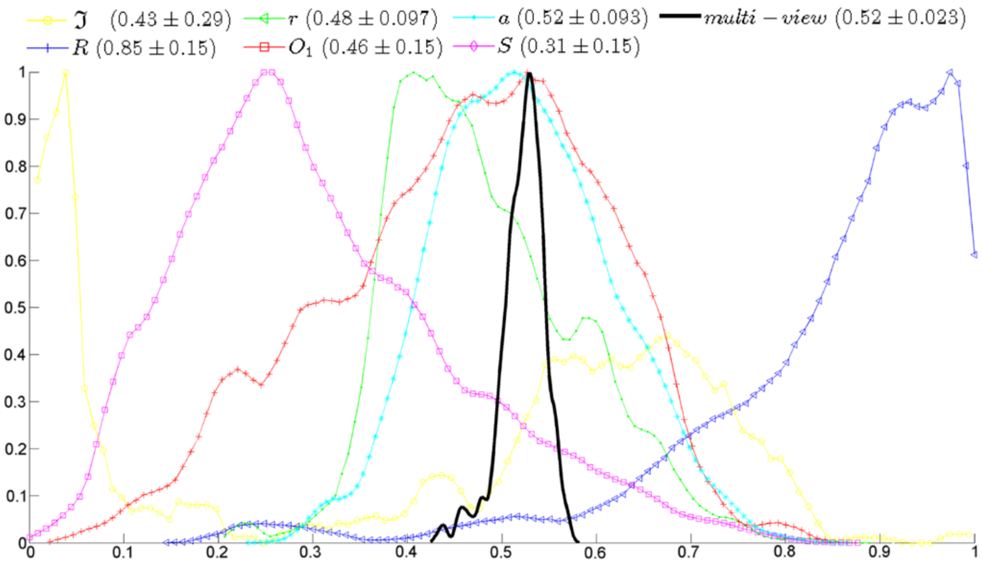


Fig. 7 Distribution of all skin pixel values in the data set for different color models. Mean and standard deviation for each channel are listed in the legend

Table 10 Performance of different detection algorithms on Caltech face database. Bold values indicate maximum performance

	$\hat{g}$	Detection accuracy	Detection rate	F
RGB based method (Fleck et al. 1996; Kovac et al. 2003)	**0.640 ± 0.19**	0.694 ± 0.20	**0.884 ± 0.17**	**0.761 ± 0.17**
CbCr based method (Chai and Ngan 1999)	0.259 ± 0.18	0.309 ± 0.21	0.548 ± 0.31	0.379 ± 0.23
HS based method (Sobottka and Pitas 1998)	0.443 ± 0.21	0.514 ± 0.21	0.807 ± 0.28	0.585 ± 0.21
RGB Statistical (Jones and Rehg 2002)	0.510 ± 0.23	0.635 ± 0.23	0.723 ± 0.28	0.643 ± 0.22
Min. Variance (Jacobs 1995)	0.189 ± 0.03	0.195 ± 0.03	0.190 ± 0.02	0.318 ± 0.05
Single-view Fusion (Stokman and Gevers 2007)	0.314 ± 0.24	0.365 ± 0.26	0.636 ± 0.34	0.430 ± 0.27
Multi-view[a]	0.410 ± 0.23	0.703 ± 0.18	0.497 ± 0.20	0.550 ± 0.15
Multi-view (our method)	0.589 ± 0.18	**0.756 ± 0.22**	0.718 ± 0.11	0.713 ± 0.17

[a] Without color model selection

Table 11 Wilcoxon test for the skin detection experiment. A positive value indicates that our method outperforms the others. Negative values indicate that our method does not perform significantly better. Bold values indicate when the proposed method outperforms the others

		RGB based	*CbCr* based	*HS* based	*RGB* Statistical	Minimum Variance	Single-view	Multi-view[a]
Multi-view	$\hat{g}$	−1	**1**	**1**	**1**	**1**	**1**	**1**
	DA	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	DR	−1	**1**	−1	−1	**1**	**1**	**1**
	F	−1	**1**	**1**	**1**	**1**	**1**	**1**

[a] Without color model selection

semble is estimated considering all the pixels from the training set to be equally relevant. However, when sequential data is considered, recent observations are more likely to occur than past ones. Thus, the optimal ensemble should be constructed emphasizing on current lighting variations rather than distant ones. Therefore, in this experiment, a sequence of more than 800 images is considered to analyze the dynamic nature of observations. This video sequence is recorded using an on-board camera. The aim of the experiment is to detect the (not occluded) road in front of a moving vehicle using a color camera. The images used include different backgrounds, the occurrence of occluding and cluttered objects (vehicles) and different road appearances under varying illumination changes.

The training set consists of 15 different road patches which are manually selected from 15 different (randomly) selected images. The selection process avoids successive image indexes. These patches contain different illumination (i.e., shadows and highlights) and they represent less than 0.053% of the total amount of road pixels within the sequence. The selection of the most suitable color models is executed by the PCA procedure described in Sect. 4. The obtained weights for the ensemble are listed in Table 7 and shows a dominant weight for the invariant color model corresponding to an achromatic surface independent of illumination changes (e.g., sun casts and shadows) i.e., roads.

Furthermore, in this experiment, the sequential nature of the data is also considered. Thus, once the optimal ensemble for the road is computed, it is adapted considering only images close in time. That is, the procedure described in Sect. 3 is used to estimate the input parameters ($E[\xi_1], \ldots, E[\xi_N]$ and Σ) to the optimization process. To estimate them, a temporal buffer is used considering the central value of the detected road in each frame for each selected color model (Fig. 8). Hence, the assumption is that the correlation between color models holds over time. To avoid possible outliers (false positives in the current result), robust statistics are used. The decay factor λ (see (12)) is empirically fixed to 0.5 as suggested by the results from the synthetic experiment in Sect. 3.1 (Table 2). Then, the optimal ensemble is recomputed at each frame considering these new values of $E[\xi_1], \ldots, E[\xi_N]$ and Σ.

To evaluate the improvement in performance when temporal information is taken into account, the error between the expected value of the road and the current value (D as in (14)) for two different updating techniques is considered (Fig. 9). The updating techniques are sample and hold, and *EWMA* or adaptive. The former uses a fixed optimal ensemble estimated using training samples over all the image sequence. The latter uses a decay factor ($\lambda = 0.5$) to update the optimal ensemble accordingly to new data available. As shown in Fig. 9, the error is significantly lower when the ensemble is updated over time. That is, if the ensemble is adapted considering new data available, then the road data distribution is modified accordingly to the new lighting conditions leading to more accurate results. However, using a fixed ensemble (sample and hold), the variations due to unobserved (not in the training set) lighting conditions or road appearances lead to shifted road data distributions. Further, the analysis of the tracking error (ψ) and historical *SR* (S_h) for both methods (Table 12) suggests that the adaptive method has a better performance in terms of following the road central value over all the images in the database.

For comparison, the video sequence is processed using three state-of-the-art methods. The first algorithm is the *HSI* road detection (RD) algorithm proposed in (Sotelo et al. 2004) and used in (Rotaru and Graf 2008). The *HSI* color space is used to process generic outdoor scenes under varying illumination (Ikonomakis et al. 2000; Sigal et al. 2004). The second algorithm is the illuminant-invariant algorithm presented in (Álvarez et al. 2008). The third algorithm is based on $2D$ histograms in rg space (Tan et al. 2006). Further, the two fusion methods proposed in (Jacobs 1995) and (Stokman and Gevers 2007) are considered. Finally, three different instances of the proposed method are considered: sample and hold without color model selection, sample and hold method using color model selection, and over time adaptive method using color model selection. Note that the *HSI* and illuminant-invariant algorithms are based on a frame-by-frame procedure. Further, these algorithms require various parameter settings. For fair comparison, a brute force approach is applied. In this way, a set of images is processed and evaluated using all possible values within the range of each parameter. The optimal set of parameter values is the one which maximizes the average performance. All algorithms (which need training) are trained using the same road pixels. Finally, all these state-of-the-art algorithms consider that the lowest part of the image corresponds to the road and that it is about 4 meters away from the vehicle. Under this consideration, only detected results which are connected with a set of seeds placed at the bottom part of the image are retrieved as road pixels. The same set of seeds is used for all the methods in the experiments.

Table 12 Tracking error (ψ) and historical *SR* (S_h) for the road experiment. The lower ψ the better performance whereas the higher S_h the better performance

	ψ	S_h
Sample and hold	0.002212	0.001
EWMA ($\lambda = 0.5$)	**0.000257**	**8.331**

The performance of all algorithms is outlined in Table 13. Various detection results of our method using temporal adaptation are shown in Fig. 10. Further, the results of the Wilcoxon test are shown in Table 14. From the results, it can be concluded that when the proposed method is adapted over time it performs significantly better than the others except for the detection accuracy for *HSI* method and the non-adapted method. Results provided by our method are slightly over-detected compared to those provided by these two methods. However, regarding the overall performance (quality and effectiveness), the proposed method performs best. This means that the proposed algorithm achieves a higher trade-off between invariance (detection rate) and discriminative power (detection accuracy).

The results reveal that the method produces false negatives (undetected road pixels) when highlights or lane markings are present. Further, the algorithm takes a few images to recover when an abundance of false positives are present. Hence, when the input data to estimate the ensemble is biased then the performance drops off. This could be improved by adding more constraints (such as unimodality test) to the new data available. Furthermore, the performance may be improved by clustering detected road pixels to distinguish different lighting conditions in the same frame.

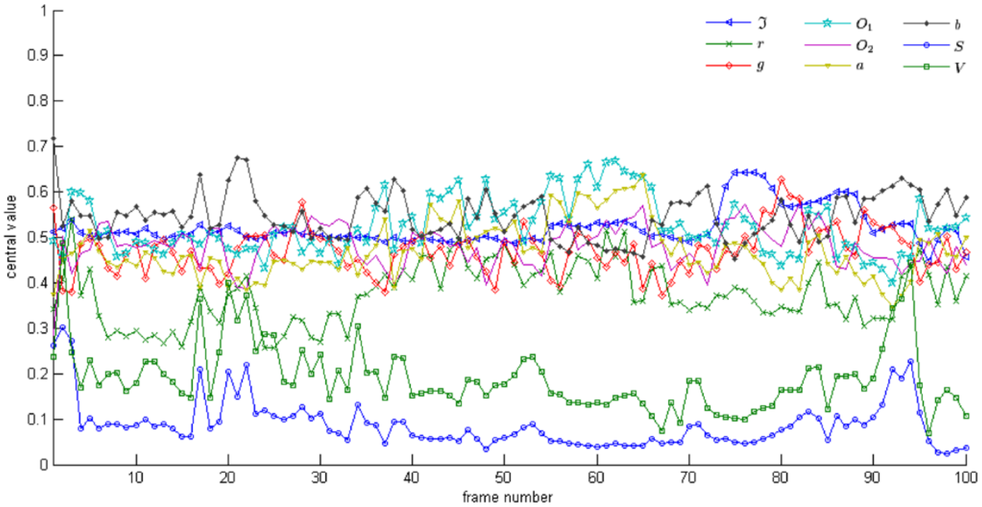


Fig. 8 Central values of observations are estimated using robust statistics on results at each frame. These values are used to recompute the optimal ensemble

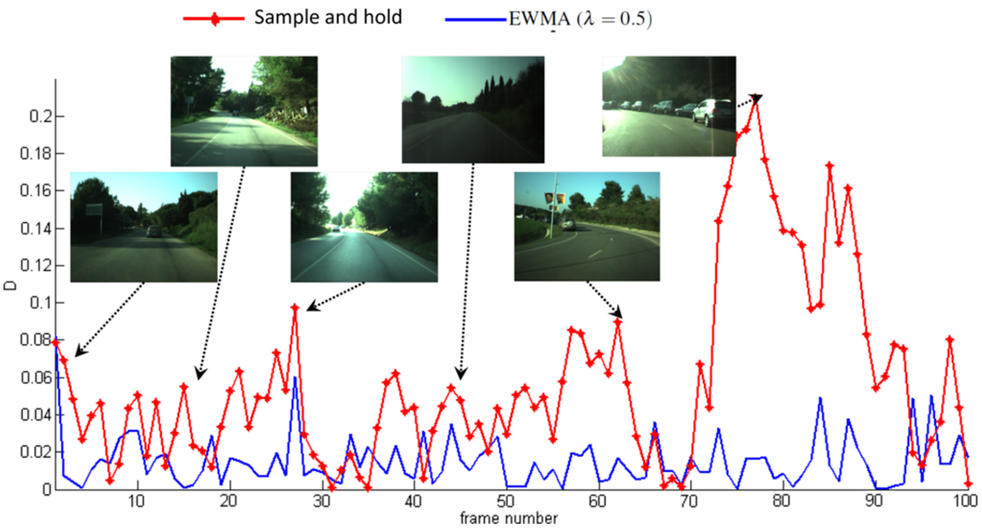


Fig. 9 Comparison between errors in the expected road value at each frame (D in Eq. (14)). For clarity reasons 1 every 100 frames are selected from the original video sequence). The error is higher when unobserved lighting conditions appear

Table 13 Performance of different detection algorithms on road database. Bold values indicate the maximum performance

	$\hat{g}$	Detection Accuracy	Detection Rate	F
HSI based RD (Sotelo et al. 2004)	0.673 ± 0.12	0.927 ± 0.12	0.729 ± 0.15	0.798 ± 0.09
Invariant RD (Álvarez et al. 2008)	0.798 ± 0.13	0.901 ± 0.15	0.866 ± 0.10	0.870 ± 0.10
rg model based (Tan et al. 2006)	0.272 ± 0.19	0.770 ± 0.23	0.410 ± 0.34	0.391 ± 0.29
Min. Variance (Jacobs 1995)	0.137 ± 0.22	0.237 ± 0.30	0.193 ± 0.31	0.187 ± 0.28
Single-view Fusion (Stokman and Gevers 2007)	0.680 ± 0.14	0.936 ± 0.02	0.716 ± 0.15	0.801 ± 0.10
Multi-view (our method)[a]	0.801 ± 0.36	0.714 ± 0.10	0.826 ± 0.05	0.746 ± 0.07
Multi-view (our method)[b]	0.810 ± 0.09	$\mathbf{0.976 \pm 0.04}$	0.828 ± 0.09	0.893 ± 0.05
Multi-view (our method)[c]	$\mathbf{0.915 \pm 0.06}$	0.963 ± 0.05	$\mathbf{0.949 \pm 0.05}$	$\mathbf{0.954 \pm 0.03}$

[a] Without color model selection

[b] Without temporal adaptation

[c] With temporal adaptation

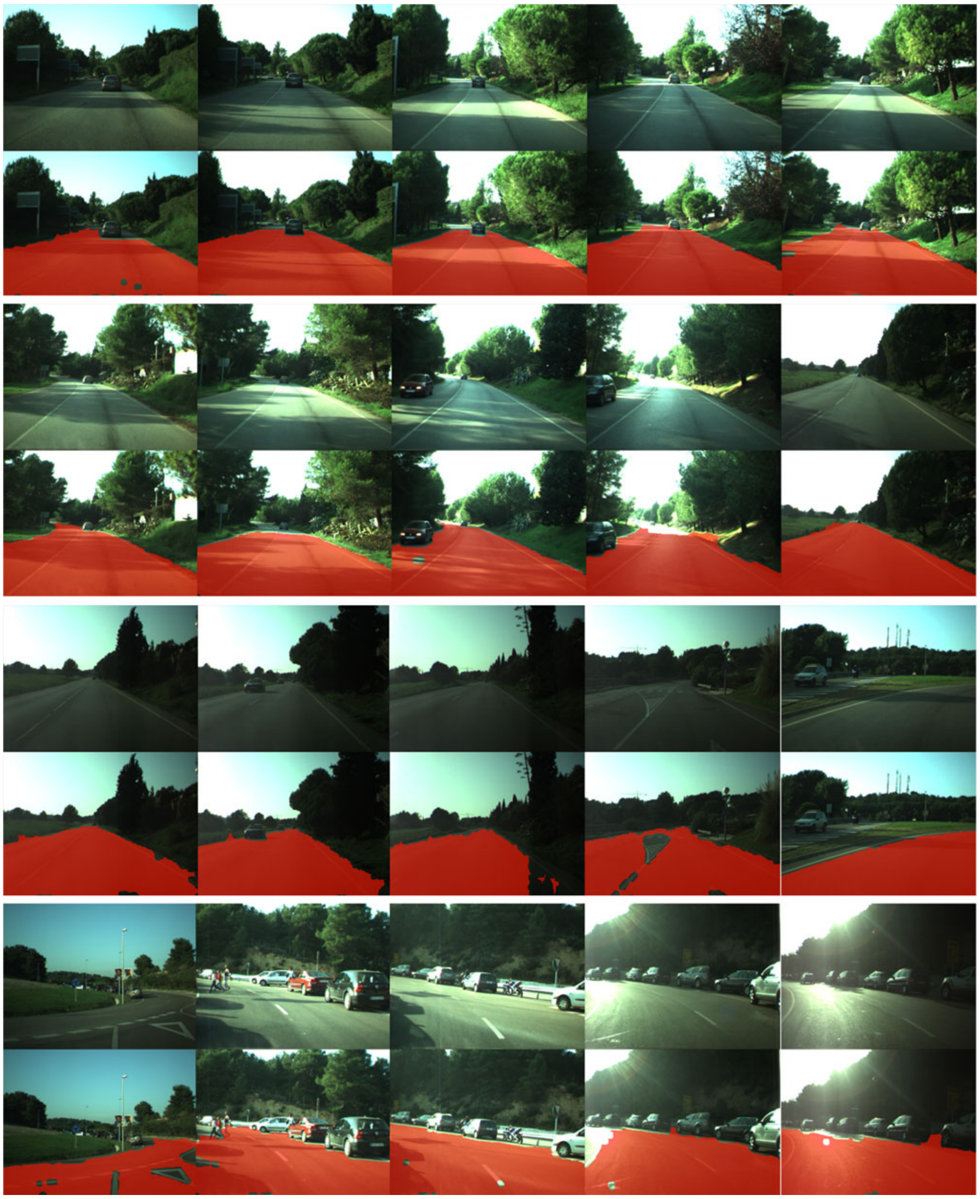


Fig. 10 Results of the proposed algorithm to detect roads

Table 14 Wilcoxon test for the road detection experiment. Positive values indicate that the proposed method performs significantly better. Negative values indicate that our method does not outperform the others. Bold values indicate when our method outperforms the others

		HSI RD	Invariant RD	*rg* model based	Minimum Variance	Single-view	Multi-view[a]	Multi-view[b]
Multi-view temporal adaptation	$\hat{g}$	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	DA	-1	**1**	**1**	**1**	**1**	**1**	-1
	DR	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	F	**1**	**1**	**1**	**1**	**1**	**1**	**1**

[a] Without color model selection

[b] Without temporal adaptation

6 Conclusions

In this paper, photometric invariance has been derived by learning from color models to obtain diversified color invariant ensembles using only positive examples. A method for combining color models is proposed to provide a multi-view approach to minimize the estimation error. In this way, the method is robust to data uncertainty and produces properly diversified color invariant ensembles. Further, the proposed method is extended to deal with temporal data by predicting the evolution of observations over time.

Experiments are conducted to validate the method. From these experiments it is concluded that our method is robust against variations in imaging conditions and is not restricted to a certain reflection model. Further, the method performs similar or outperforms state-of-the-art detection techniques in the field of object, skin and road recognition. Considering sequential data, the proposed method that is extended to deal with future observations outperforms the other methods.

Acknowledgements This work was supported by the Spanish Ministry of Education and Science under project TRA2007-62526/AUT and research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

References

Álvarez, J. M., López, A. M., & Baldrich, R. Illuminant-invarariant model-based road segmentation. In *Proceedings of the 2008 IEEE international vehicles symposium (IV'08)*, Eindhoven, The Netherlands.

Best, P. (1998). *Implementing value at risk*. New York: Wiley.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: a survey and categorisation.

Chai, D., & Ngan, K. (1999). Face segmentation using skin-color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology*, *9*(4), 551–564.

Dowd, K. (1998). *Beyond value at risk: the new science of risk management*. New York: Wiley.

Finlayson, G. D., Drew, M. S., & Lu, C. (2004). Intrinsic images by entropy minimization. In *Proceedings of the European conference on computer vision (ECCV)* (Vol. 3, pp. 582–595).

Finlayson, G., Hordley, S., Lu, C., & Drew, M. (2006). On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(1).

Fleck, M. M., Forsyth, D. A., & Bregler, C. (1996). Finding naked people. In *Proceedings of the European conference on computer vision (ECCV)* (Vol. 3, pp. 593–602). Berlin: Springer

Geusebroek, J. M., Burghouts, G. J., & Smeulders, A. W. M. (2005). The Amsterdam library of object images. *International Journal Computer Vision*, *61*(1), 103–112.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70–84.

Ikonomakis, N., Plataniotis, K., & Venetsanopoulos, A. (2000). Color image segmentation for multimedia applications. *Journal of Intelligent Robotics Systems*, *28*(1–2).

Jacobs, R. A. (1995). Methods for combining experts' probability assessments. *Neural Computing*, *7*(5), 867–888.

Jolliffe, I. T. (2002). *Springer series in statistics*. *Principal component analysis* (2nd ed.). Berlin: Springer.

Jones, M. J., & Rehg, J. M. (2002). Statistical color models with application to skin detection. *International Journal Computer Vision*, *46*(1), 81–96.

Kender, J. (2005). *Saturation, hue and normalized color: calculation, digitation effects, and use* (Tech. Rep. CMU-RI-TR-05-40). Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Kittler, J., Hatef, M., Duin, R., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239.

Kovac, J., Peer, P., & Solina, F. (2003). Human skin color clustering for face detection. In *International conference on computer as a tool (EUROCON)*.

Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. New York: Wiley-Interscience.

Markowitz, H. M. (1959). *Portfolio selection: efficient diversification of investments*. New York: Wiley.

Melville, P., & Mooney, R. J. (2005). Creating diversity in ensembles using artificial data. *Information Fusion*, *6*(3), 1553–1563.

Michaud, R. O., & Michaud, R. (2008). Estimation error and portfolio optimization: a resampling solution. *Journal of Investment Management*, *6*(1), 8–28.

Michaud, R. O. (1998). *Efficient asset management: a practical guide to stock portfolio optimization and asset allocation*. Oxford: Oxford University Press.

Rasmussen, M. (2003). *Quantitative portfolio optimisation, asset allocation and risk management (Finance and capital markets)*. Basingstoke: Palgrave Macmillan.

Ridler, T., & Calvard, S. (1978). Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man and Cybernetics*, *8*(8), 630–632.

Rotaru, C., Graf, T., & Zhang, J. (2008). Color image segmentation in hsi space for automotive applications. *Journal of Real-Time Image Processing*, 1164–1173.

Scherer, B. (2002). *Portfolio construction and risk budgeting* (Chap. 4). London: Rosk Books.

Sharpe, W. (1994). The sharpe ratio. *Journal of Portfolio Management*, *21*, 49–58.

Sigal, L., Sclaroff, S., & Athitsos, V. (2004). Skin color-based video segmentation under time-varying illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(7), 862–877.

Sobottka, K., & Pitas, I. (1998). A novel method for automatic face segmentation, facial feature extraction and tracking. *Signal Processing: Image Communication*, *12*(3), 263–281.

Sotelo, M., Rodriguez, F., Magdalena, L., Bergasa, L., & Boquete, L. (2004). A color vision-based lane tracking system for autonomous driving in unmarked roads. *Autonomous Robots*, *16*(1).

Stokman, H., & Gevers, T. (2007). Selection and fusion of color models for image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 371–381.

Tan, C., Hong, T., Chang, T., & Shneier, M. (2006). Color model-based real-time learning for road following. In *Proceedings of the IEEE international conference on intelligent transport systems* (pp. 939–944).

Tax, D. M. J., & Duin, R. P. W. (2002). Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, *2*, 155–173.

Tse, Y. K. (1991). Stock returns volatility in the Tokyo stock exchange. *Japan and the World Economy*, *3*(3), 285–298.

Usmen, N. M. H. (2003). Resampled frontiers versus diffuse Bayes: an experiment. *Journal of Investment Management*, *1*(4), 1–17.

van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2008). Evaluation of color descriptors for object and scene recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 453–464).

Weber, M. (1999). *The Caltech frontal face dataset*. California Inst. of Tech., USA. http://www.vision.caltech.edu/html-files/archive.html. Accessed 1 March 2010.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Wyszecki, G., & Stiles, W. (1982). *Color science: concepts and methods, quantitative data and formulae* (2nd ed.). New York: Wiley.