The International Conference on Computer Vision (ICCV03)
October 13-16, 2003, Nice, France

On the Interaction between Object Recognition and Colour Constancy

Štěpán Obdržálek[1]
xobdrzal@fel.cvut.cz

Jiří Matas[1,2]
matas@cmp.felk.cvut.cz

Ondřej Chum[1]
chum@cmp.felk.cvut.cz

[1]Center for Machine Perception, Czech Technical University, Prague, 120 35, CZ
[2]Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK

Abstract

In this paper we investigate some aspects of the interaction between colour constancy and object recognition. We demonstrate that even under severe changes of illumination, many objects are reliably recognised if relying only on geometry and on invariant representation of local colour appearance. We feel that colour constancy as a preprocessing step of an object recognition algorithm is important only in cases when colour is major (or the only available) clue for object discrimination.

We also show that successful object recognition allows for ”colour constancy by recognition” – an approach where the global photometric transformation is estimated from locally corresponding image patches.

1. Introduction

In this paper we investigate some aspects of the interaction between colour constancy and object recognition. Colour constancy is a classical problem that has been recently connected to object recognition [11, 5, 2]. In [5], Funt et al. propose to judge the quality of colour constancy algorithms by their impact on recognition rates. The question ”Is colour constancy good enough (for object recognition)” is posed. For histogram intersection as the recognition method and a wide range of colour constancy algorithms the answer is *negative*, i. e. none of the tested colour constancy algorithms is ”good enough”.

We revisit the issue and show that if a recognition method relies mainly on geometry and representation of local colour appearance invariant to affine transformation of colour components (equivalent to a diagonal colour constancy model [3] with an offset term), object recognition can be successful even under severe and unknown change of illumination. This is experimentally demonstrated on a public dataset from the Simon Fraser University, that has been previously used in colour constancy experiments [1, 2].

Successful recognition insensitive to illumination allows us to consider the intuitive approach of ”colour constancy by recognition”. We show that a straightforward approach which estimates the colour transformation from local correspondences established in the recognition step is more precise than the best standard (global, correspondence-less) colour constancy method. The precision of ”colour constancy by recognition” is measured by the distance (in the chromatic plane) of the white point under canonical illumination and the transformed white point of the image under the unknown illumination. The achieved precision is approximately three times higher than that of Barnard et al [2].

The result has to be interpreted carefully. Clearly, the presence of a known object in the scene is a restrictive assumption. Colour constancy is often required in scenes without known object, e.g. as a part of a white balance module of a camera. The message is rather that if a known object is in the scene, much better results of colour constancy can be expected, if the object is recognised. It seems that two different classes of colour constancy algorithms might be distinguished: those relying on global or statistical properties and those attempting to recognise object or object classes (hair, skin) and use constraints on scene illumination imposed by observed colours of known surfaces. Unlike the former, the latter colour constancy algorithms are able to deal with non-uniform illumination. In a synthetic experiments, we show that it is possible to partition the image according the illuminant.

The rest of the paper is structured as follows. In Section 2, we review a recognition method based on the concept of local affine frames. Locally, the image is photometrically normalised to compensate for affine transformation of each colour channel. The normalisation is detailed in Section 3 together with the matching strategy for establishing local image-to-image correspondences. An approach to ”colour constancy by recognition” is proposed in Section 4. A full affine model for the global photometric transformation is adopted. Two experiments are described in Section 5. First, the recognition performance of the local affine frame method is tested in changing lighting conditions. In second experiment, a colour transformation to a canonical illumination is estimated and its precision measured. The paper is concluded in Section 6.

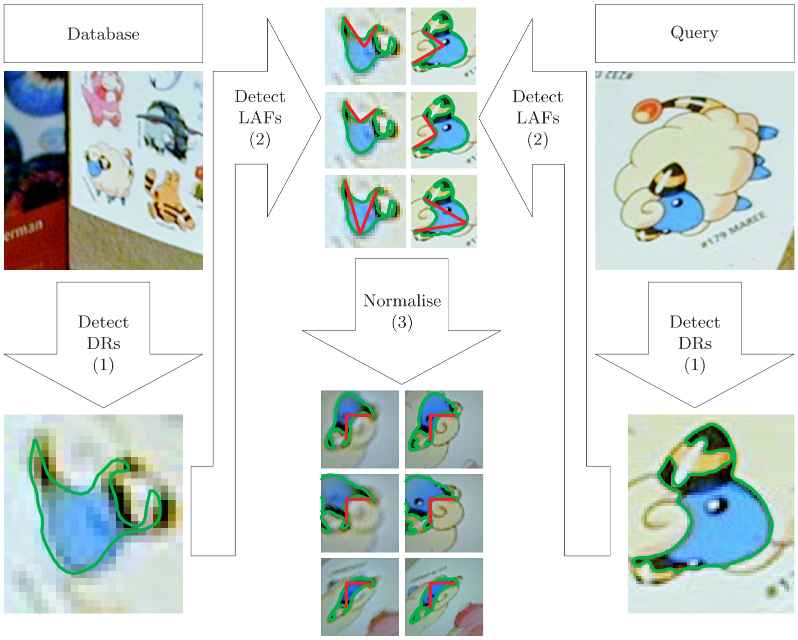


Figure 1: Local affine invariant image descriptors. Structure of computation.

2. Overview of the Matching Process

The outline of the method is following (the first three steps are visualised in Fig. 1):

1. For every database and query image compute distinguished regions (DRs).
2. Construct local affine frames (LAFs) on the regions.
3. Generate intensity representations of local image patches normalised according to the local affine frames. and photometrically normalise the intensity representation.
4. Establish correspondences between frames of query and database images, by computing the euclidean distance between the local image intensities, and by finding the nearest match.
5. An estimate of the match score is based on the number and quality of the established local correspondences.
6. If an object is recognised, global photometric transformation is estimated between database and query images

In the rest of this Section we briefly introduce the concepts of the first two steps, the distinguished regions and the local affine frames. Remaining steps are discussed in the following sections.

Distinguished Regions (DRs) are image elements (subsets of image pixels), that posses some distinguishing property that allows their repeated and stable detection over a range of image formation conditions. In this work we exploit the distinguished regions introduced in [7], the *Maximally Stable Extremal Regions* (MSERs). MSERs are image elements detected by local thresholding of a greyscale image and are stable under monotonic transformations of the grey-values. The transformation between RGB image and the greyscale image used for region detection can be arbitrary. We use the intensity component of the colour image as it is stable under a wide range of illumination changes. But generally, severe change in illumination can re-order the intensities between the images (the grey-values transformation may not be monotonic). MSERs will then fail.

For further reference on MSERs see [7] which includes a formal definition and a detailed description of the extraction algorithm.

Local affine frames (LAFs, local object-centered coordinate systems) allow normalisation of image patches into a canonical frame, and enable direct comparison of photometrically normalised intensity values, eliminating the need for invariants. For every distinguished region, multiple frames are computed. The actual number of the frames depends on the region's complexity. While simple elliptical regions have no stable frames detected, regions of complex non-convex shape may have tens of frames associated. Robustness of our approach is thus achieved by 1. selecting only stable frames and 2. employing multiple processes for frame computation. A detailed description of the local affine frame constructions is given in [8], [9] and [10].

3 Photometric and Geometric Normalisation

Each image is represented by a set of local measurements. Once local affine frames are established, there is no need for geometrically invariant descriptors of local appearance. Any measurement taken relative to the frame is affine invariant.

Geometry. The affine transformation between the canonical frame with origin $O = (0,0)^T$ and basis vectors $e_1 = (1,0)^T$ and $e_2 = (0,1)^T$ and an established frame F is described in homogenous coordinates by a 3 by 3 matrix

$$\mathbf{A}_F = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{pmatrix}.$$

The image patch (defined in terms of the affine frame) where the local measurements are taken from is referred to as a measurement region (MR). The choice of MR shape and size is arbitrary. Larger MRs have higher discriminative potential, but are more likely to cover part of an object that is not locally planar. Our choice is to use a square MR centered around a detected LAF, specifically a region spanning $\langle -2, 3\rangle \times \langle -2, 3\rangle$ in the frame coordinate system. Transformed to the image coordinate system, the measurement region of a frame F becomes a parallelogram with corners at (in homogenous coordinates):

$$c_1 = \mathbf{A}_F \begin{pmatrix} -2 \\ -2 \\ 1 \end{pmatrix}, \quad c_2 = \mathbf{A}_F \begin{pmatrix} -2 \\ 3 \\ 1 \end{pmatrix},$$

$$c_3 = \mathbf{A}_F \begin{pmatrix} 3 \\ -2 \\ 1 \end{pmatrix}, \quad c_4 = \mathbf{A}_F \begin{pmatrix} 3 \\ 3 \\ 1 \end{pmatrix},$$

Photometry. For the process of establishing local correspondences we utilise a simple photometric model. We assume a linear camera (ie. a camera without gamma-correction). Specular reflections are ignored. The combined effect of different scene illumination and camera and digitiser settings (gain, shutter speed, aperture) is modelled by affine transformations of individual colour channels. The photometric transformation between two corresponding patches I and I' is considered in the form:

$$\begin{pmatrix} r' \\ g' \\ b' \end{pmatrix} = \begin{pmatrix} m_r & 0 & 0 \\ 0 & m_g & 0 \\ 0 & 0 & m_b \end{pmatrix} \begin{pmatrix} r \\ g \\ b \end{pmatrix} + \begin{pmatrix} n_r \\ n_g \\ n_b \end{pmatrix}$$

The constants m_r, n_r, m_g, n_g, m_b, n_b differ for individual correspondences. This model would agree with the monochromatic reflectance model [6] in the case of narrow band sensor. It can be viewed as an affine extension of the diagonal model, that has been shown by Finlayson to be sufficient in common circumstances [4][1].

To represent the patch invariantly to photometric transformations, intensities are transformed into a canonical form. The intensities of individual colour channels are affinely transformed to have zero mean and unit variance.

Let us summarize the **Normalisation Procedure** of a local patch:

1. Establish a local affine frame F.

2. Compute the affine transformation $\mathbf{A}_F$ between the canonical coordinate system and F.

3. Express the intensities of the F's measurement region in the canonical coordinate system
   $I'(\mathbf{x}) = I(\mathbf{A}_F\mathbf{x}), \quad \mathbf{x} \in$ MR with some discretisation.

4. Apply the photometric normalisation
   $\hat{I}'(\mathbf{x}) = (I'(\mathbf{x}) - \mu)/\sigma, \quad \mathbf{x} \in$ MR
   where μ is the mean and σ is the standard deviation of I' over the MR.

The twelve normalisation parameters ($a_1 \dots a_6$, m_r, n_r, m_g, n_g, m_b, n_b) are stored along with the normalised intensity measurement. When considering a pair of patches for a correspondence, these twelve parameters are combined to provide the local transformation (both geometric and photometric) between the images.

The correspondences are formed by evaluating the correlation coefficient between discretised representations of the normalised measurement regions. Tentative correspondences are formed if the coefficient is above a predefined threshold. In a second step, the subsets of geometrically and photometrically consistent tentative correspondences are found. Examples of pairs of corresponding patches (MRs) are depicted in Figure 2.

4. Estimating the Photometric Transformation

Local measurements are constructed with invariance to diagonal (or affine extension of diagonal) photometric transformations, as described in Section 3. At local scale, such a simple photometric model is sufficient to establish correspondences. Global colour transformation is computed after the correspondences are found, using full affine model. By

[1] At least in conjunction with sensor sharpening [3]

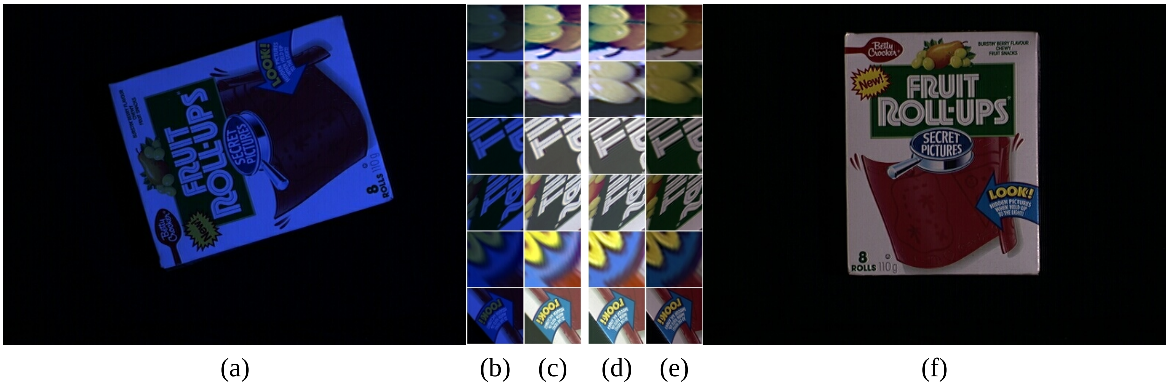


Figure 2: Normalised local correspondences. (a), (f): Query and Database images, (b), (e): Examples of geometrically normalised MRs (measurement regions), (c), (d): Photometrically normalised MRs

considering only the image regions that were put into correspondence, the global transformation is found independently of any background clutter or occluding objects.

Establishing Pixel-to-Pixel Correspondences. Every established correspondence locally maps a pair of regions. Assuming that local geometric deformations are sufficiently well approximated by 2D affine transformations, pixel correspondences are obtained by sampling the images with respect to the local coordinate systems of corresponding LAFs. This can be interpreted as a regular sampling of the geometrically normalised MRs depicted in Figure 2 (b) and (e). In our implementation, we sample the MRs on a regular 6×6 grid, obtaining thus 36 pixel-correspondences per every frame-correspondence. For a typical object, the number of pixel-correspondences is in the order of thousands.

Computing the Photometrical Transformation. With thousands of corresponding pixels available, the global photometric transformation can be calculated in a more complicated form than the diagonal, without the risk of overfitting. We compute the transformation in its affine form, i.e.

$$\begin{pmatrix} r' \\ g' \\ b' \end{pmatrix} = \begin{pmatrix} m_1 & m_2 & m_3 \\ m_4 & m_5 & m_6 \\ m_7 & m_8 & m_9 \end{pmatrix} \begin{pmatrix} r \\ g \\ b \end{pmatrix} + \begin{pmatrix} n_r \\ n_g \\ n_b \end{pmatrix}$$

The transformation coefficients are obtained by least squares fitting, i.e. the sum of square differences between transformed colours of query pixels and colours of corresponding database pixels is minimised.

5. Experiments

Dataset. The experiments were realised on a publicly available dataset published by Barnard [1]. The dataset contains images of 20 different objects, every object is taken under 11 illuminants. The total number of images in the dataset is thus 220. The illuminants were chosen to cover the range

Method	Recognition rate
LAFs	89.1 %
Hist. Intersection, no CC	42.3%
Hist. Intersection, manual CC	87.7%
Hist. Intersection, best CC	80.9%
Hist. Intersection, worst CC	15.5%

Table 1: Summary of the recognition experiment

Illuminant	Recogn. rate	WP error
ph-ulm	17/20, 85%	0.015
solux-3500+3202	19/20, 95%	0.011
solux-3500	19/20, 95%	0.006
solux-4100+3202	17/20, 85%	0.013
solux-4100	20/20, 100%	0.008
solux-4700+3202	12/20, 60%	0.021
solux-4700	19/20, 95%	0.012
syl-50MR16Q+3202	18/20, 90%	0.009
syl-50MR16Q	20/20, 100%	–
syl-cwf	16/20, 80%	0.010
syl-wwf	19/20, 95%	0.013
average	89%	0.012
best method in [2]	81%	0.038

Table 2: Recognition rate and illuminant colour estimation

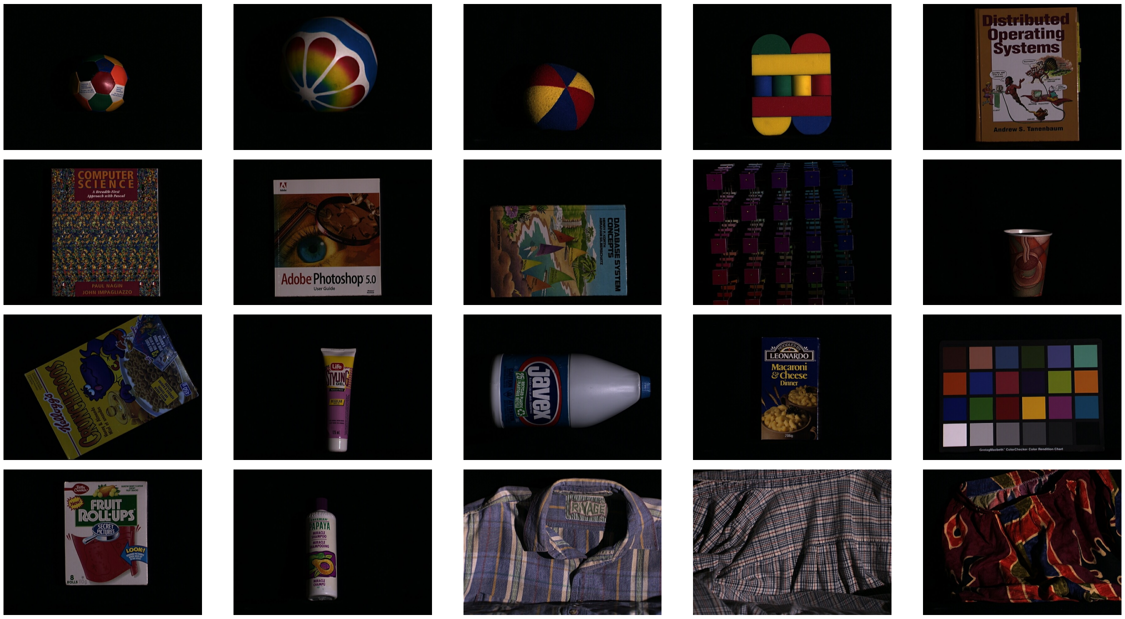


Figure 3: All 20 database images.

of common illumination conditions. For each image, chromaticity of the white point is provided. It was obtained by temporarily placing a sheet of white paper in the scene.

The object recognition task is simplified by the fact that the objects are placed on black background (ie. there is no background clutter, the objects can be segmented out) and the objects are unoccluded. However, the objects were taken in different poses. In some cases, only different parts of their surfaces are visible. All the database objects are shown in Figure 3.

Experimental Protocol. The training database (the set of known images) contains a single image of every object. We have used the images taken under illuminant 'syl-50MR16Q'. To follow the experimental setup from [2], all 220 images are used as queries, ie. the set of queries contains also the database images. Every query image is matched against every database image. As there are no images of non-database objects, the database image with the highest score is always selected (forced match).

We manually selected those query – database image pairs where the object was successfully recognised. The global colour transformation between the query and the database images was estimated based on the photometric transformations computed from corresponding regions, as described in Section 4. The transformation was estimated as full affine, i.e. with 12 degrees of freedom.

The query-to-database photometric transformation can not be used to estimate the colour of the illuminant (ie. the white point) since image taken under "white light" are not part of the database. The precision of the estimated global photometric transformation is verified by transforming the provided white paper colour of the query image. Ideally, the transformed colour should be equal to the white paper colour of the matched database image. As it is not, the precision of the estimate is measured by computing the euclidean distance between chromaticities of the transformed query white point and the database white point.

Results. Results of the recognition experiments are sumarised in Table 1. Our method (LAFs) is compared to results published in [2]. In [2], query images are first adjusted by one of a rather exhaustive set of 23 colour constancy algorithms. The matching is then done by histogram intersection method on the adjusted images.

The first row of Table 1 shows the recognition rate of our method, second row of the histogram intersection method without any colour constancy being applied. The third row shows results for manual colour constancy, where the query images were transformed so that the manually measured white points match. The remaining two rows report results for the best (non-diagonal, coefficient-rule) and the worst (color-in-perspective) of the 23 colour constancy algorithms. Our recognition performance is superior to any of the results presented in [2].

Table 2 shows how individual illuminants affect recognition rate of the LAF method. There is no significant difference in the performance, except for the 'solux-4700+3202' illuminant (4700K incandescent light plus a blue filter). The recognition failures here are not due to the illuminant colour, but due to the low intensity of the images captured under this light. The third column of Table 2 shows the

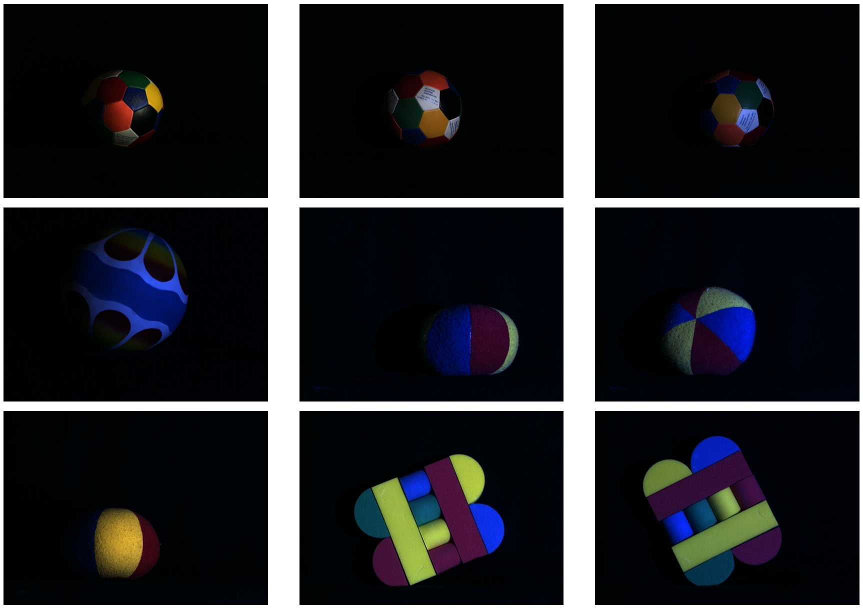
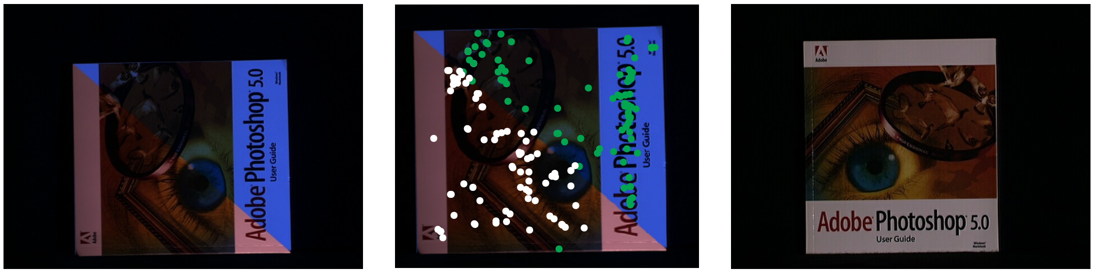


Figure 4: Examples of recognition failures. The objects are not recognised due to different pose, not due to illumination.



(a) (b) (c)

Figure 5: Scene with multiple illuminants: (a) an synthetic query image, two differently illuminated halves joined, (b) found correspondences clustered by local photometric transformation, (c) corresponding database image

precision of the global photometric transformation estimation. For comparison, a white point estimation error of the best performing method from [2] is quoted. Our estimates are on average three times more precise, but note that only correctly recognised images are included. Estimation based on mismatched objects may produce arbitrary photometric transformation.

Figure 4 shows all our recognition failures in queries for the first four database objects. The query images differ from the database images not only in the illumination, but, more significantly, in the object pose. The balls are rotated so that their visual appearance is substantially different from the database images. The blocks-object was turned upside-down, producing a 'mirror' image of itself, which is not recognised by our method. Refer to Figure 3 to see the differences between database images and the unrecognised queries.

Multiple Illuminants. In a final experiment, we demonstrate that our recognition system can handle objects viewed under multiple illuminants at the same time, as can be the case when a shadow is cast over part of an object. Figure 5 (a) shows our query image, which was obtained by artificially merging two images of the object. The process of image description and matching is invariant to local illumination. Presence of multiple illuminants thus have no effect on the obtained correspondences, except for LAFs that are on the boundary of differently illuminated object areas.

Correspondences are clustered by their local photometric transformation. Each such cluster represents a global transformation caused by one of the illuminants. In Figure 5 (b) two clusters of correspondences are shown as green and white dots respectively. With a single exception, the correspondences are correctly separated according to the illuminant.

Summary. We have experimentally shown that our geometry-based object recognition method outperforms the methods described in [2], ie. the histogram intersection algorithm applied after colour constancy correction. The recognition rate of our system was almost independent of the illuminant, changes in objects' poses had a much stronger impact on the results. When an object was correctly recognised, even a straightforward least-squares algorithm was able to estimate the global photometric transformation three times more precisely than the best correspondence-less colour constancy method published in [2].

Finally, an experiment on a scene where different parts of the image are illuminated by different light sources was shown. Computing global colour transformation to a canonical illumination in such a scene is an ill-posed task. The image was however successfully recognised, partitioned according to the colour of incident light and the illumination for each part was correctly estimated by the proposed method.

6. Conclusions

In this paper we have revisited the connection between colour constancy and object recognition. We have demonstrated that for many objects a recognition method relying mainly on geometry and invariant representation of local colour appearance can be successful even under severe and unknown changes of illumination. Successful object recognition allows for ”colour constancy by recognition” – an approach where the global photometric transformation is estimated from locally corresponding image patches. In our experiments, such estimate was three times more precise than that of any global colour constancy method published.

If the known objects in the scene have strong geometric features, recognition can provide good colour constancy. On the other hand, if the objects do not have distinctive parts, or if their structure is not preserved (e.g. by non-rigid object deformation), recognition by colour or by texture becomes necessary. In this case, colour constancy can support recognition.

References

[1] Kobus Barnard. Data for computer vision and computational colour science. http://www.cs.sfu.ca/ colour/data/.

[2] Kobus Barnard, Brian Funt, and Lindsay Martin. Color constancy meets color indexing. http://vision.cs.arizona.edu/kobus/research/publications/indexing, 2000.

[3] G. D. Finlayson, M. S. Drew, and B. V. Funt. Spectral sharpening: Sensor transformations for improved color constancy. *Journal of the Optical Society of America*, 11:1553–1563, 1994.

[4] G. D. Finlayson, M.S. Drew, and B.Funt. Color constancy: Generalized diagonal transforms suffice. *Journal of the Optical Society of America*, 11:3011–3019, 1994.

[5] Brian Funt, Kobus Barnard, and Lindsay Martin. Is colour constancy good enough? In *5th European Conference on Computer Vision*, pages 445–459, 1998.

[6] G. Healey. Using color for geometry-insensitive segmentation. *Journal of the Optical Society of America*, 6:86–103, June 1989.

[7] Jiří Matas, Ondřej Chum, Martin Urban, and Tomáš Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Paul L. Rosin and David Marshall, editors, *Proceedings of the British Machine Vision Conference*, volume 1, pages 384–393, London, UK, September 2002. BMVA.

[8] Jiří Matas, Štěpán Obdržálek, and Ondřej Chum. Local affine frames for wide-baseline stereo. In *ICPR02*, August 2002.

[9] Štěpán Obdržálek and Jiří Matas. Local affine frames for image retrieval. In *The Challenge of Image and Video Retrieval (CIVR2002)*, July 2002.

[10] Štěpán Obdržálek and Jiří Matas. Object recognition using local affine frames on distinguished regions. In *The British Machine Vision Conference (BMVC02)*, September 2002.

[11] M. Swain and D. Ballard. Color indexing. In *International Journal of Computer Vision, vol. 7, no. 1*, pages 11–32, 1991.