Signal Processing 93 (2013) 1485–1495

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Learning colours from textures by sparse manifold embedding

Jun Li*, Wei Bian, Dacheng Tao, Chengqi Zhang

Centre for Quantum Computation & Intelligent Systems, University of Technology, Sydney

ARTICLE INFO

Article history:
Received 21 January 2012
Received in revised form
17 July 2012
Accepted 12 August 2012
Available online 21 August 2012

Keywords:
Sparse regression
Manifold learning
Colourisation

ABSTRACT

The capability of inferring colours from the texture (grayscale contents) of an image is useful in many application areas, when the imaging device/environment is limited. Traditional manual or limited automatic colour assignment involves intensive human effort. In this paper, we have developed a user-friendly colourisation technique, where the algorithm learns the relation between textures and colours in a user-provided example image and applies the relation to predict the colours in the target image.

The key contribution of the proposed technique is trifold. First, we have explicitly built a linear model for the texture–colour relation. Second, we have considered the global non-linear structure of the data distribution by applying the linear model *locally*; and the local area is determined automatically by sparsity constraints. Third, we have introduced semantic information to further improve the colourisation. Examples demonstrate the effectiveness of the proposed techniques. Moreover, we have conducted a subjective study, where user experience supports the superiority of our method over existing techniques.

1. Introduction

Human vision percepts the world with colours. Colours do not only make images feel more vivid to viewers, they also contains important visual clues of the image [17]. Although a modern point-and-shoot digital camera can easily capture colour images, there are circumstances where we need to recover the chromatic information in an image. For example, photography in old days was monochrome and provided only gray-scale images. Adding colours can rejuvenate these old pictures and make them more adorable as personal memoir or more accessible as archival documents for public or educational purposes. For a colour image, re-colourisation may be necessary if the white balance was poorly set when shooting the picture. In this case, a particular colour channel can be severely over- or under- exposure, and makes infeasible to adjust the white balance based on the recorded colours. A possible rescue of the picture is to keep only the luminance and re-colourise the image. Another example of the application of colourisation arises from the area of specialised imaging, where the sensors captures signals that are out of the visible spectrum of light, e.g. X-ray, MRI, near infrared images. Pseudo colours for these images make them more readily for interpretation by human experts, and can also indicates potentially interesting regions.

Consistent and efficient image colourisation is a non-trivial problem because of a basic ambiguity of the relation between luminance and colours: a pixel of a particular luminance intensity can have any colour. It seems that the brightness conveys little chromatic information and colourisation is an ill-posed problem. However, given a gray-level image, a human viewer often has a rough idea about the missing colours. This is because the intensity values at the pixels in an image are not independent, but make textures. Human recognises familiar elements from the textures and infer the corresponding colour patterns. With the help of the side information, image colourisation becomes possible. However, manual

* Corresponding author.
E-mail address: junjy007@googlemail.com (J. Li).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.08.008

implementation of pixel-wise colourisation is impractical; and the subjective guess tends to be inconsistent. A semi-automatic solution is to let a human supervisor indicate essential clues of the contents and corresponding colours in an image and to let a computer to complete the colours at each pixel [27]. This scheme brings complementary advantages from both sides of human and machine. The human user provides high level information about the contents of an image; and the machine computes the colour details at a large number of pixels.

The communication of the high level knowledge is essential for the utility of the colourisation system. Let us consider a scenario that a user identifies textures of plant leaves in an image. Then there are several choices of notifying the system about the content and the assumed colours, e.g. (i) choosing a particular greenish colour from a palette, (ii) tagging the region as "leaf" and letting the system determine the colour or (iii) specifying some example images of leaves and letting the system learn colours from the examples [38]. Method (i) is the most straightforward from the viewpoint of the system, however, it requires very tedious efforts on the users' side. Method (ii) is easy for the user, but it rises an open problem of retrieving imagery information from text descriptions, which deserves much research for its own sake. Method (iii), on the other hand, involves slightly more human effort than method (ii), but assists the system with much more information. In the paper, we are concerned with the problem of developing an effective example based colour transfer algorithm. The goal is to automatically learn the colour–texture relations in the examples and recovers the colours of the gray-scale query.

The key contribution of the work consists of three techniques advancing the previous algorithm [38] of the example-based colour transfer. First, we propose to consider the image textures as distributed on manifolds, and formulate colour prediction as a regression problem based on this assumption. Second, we propose to adopt the sparsity constraint towards automatically determining the neighbourhood size of the manifold to be considered in the regression problem. Thirdly, we propose to employ semantic information to eliminate irrelevant inputs for the system. In particular, the pixels of the images are clustered according to affinity-transform-invariant local features. The clustering makes a semantic map of the images. In predicting colours for a query, we consider only the colours from the example images, where semantic maps match that of the query.

Both objective and subjective studies have been conducted to show the effectiveness of the proposed scheme. In the objective study, it has been shown that the proposed techniques progressively improve the prediction performance. In the subjective study, the user experience of volunteers favours the proposed method over the previous technique.

The remainder of the paper is organised as follows. Section 2 provides a brief review of related techniques. Section 3 presents technical details, where the proposed algorithm is fully specified and the underlying motivation is discussed. Section 4 demonstrates the proposed colourisation technique by examples, and shows its effectiveness by reporting both the objective evaluation and users' subjective assessment. Section 5 concludes the paper with further discussion.

2. Related work

Automatic systems for assigning colours to images has been developed in the cartoon industry since long ago. Early systems needs intensive human interaction [27]. Qu et al. proposed an method to alleviate human effort [24] but the system needs an initial estimation of the colours for each components in the image, which may pose difficulties for users without corresponding expertise. Moreover, direct user interaction prevents batch processing: assigning colours to multiple gray-scale images efficiently. As we have mentioned, the example-based scheme adopted in this paper is proposed by Welsh et al. [38]. This semi-automatic scheme minimises human intervention and only requires the user to provide a reference colour image of similar contents as the interested gray-scale image. The algorithm in [38] predicts colours at each pixel of the gray-scale image by looking up colours of corresponding pixels in the reference image. The correspondence between two pixels A and B is determined by simple statistics of the luminance of pixels in respective surrounding areas of A and B. To a wider sense, the texture information at the local scale in an image and its connection to the chromatic information are also studied in [29,28,40].

This paper is based on our conference publication [21,20]. In [21], we proposed to use more textural details at a pixel to represent corresponding luminance information at the position. The structures of the texture and colour populations are represented by a linear model, which helps predict unknown colour. The model has been built based on the manifold assumption of the populations [4,26]. Later in [20], sparsity constraint is introduced in the model, which improves the stability of the algorithm with respect to the uncertainty of the local structure of the unknown manifold. In this paper, we have developed from the summarisation of [21,20]. We have introduced semantic information to help control artefacts of the prediction. A subjective study has been conducted to measure the quality of the reconstructed colours.

To make efficient use of the information in the given examples, it is essential to match an unseen sample to relevant examples in the training set, which requires meaningful measurement of the similarity between the samples. Rich studies are devoted to this problem [19,1,37,22,18]. As mentioned above, we adopt the strategy of assuming a manifold structure of the data distribution, where the low dimensional manifold helps define meaningful geometric structures in the sample space. Early research has demonstrated that the manifold assumption is suited for the feature space of image data [4]. The past decade has witnessed the emergence and flourish of tools that analyse manifold data and their applications in wide areas [34,13,14], e.g. face and biometric recognition, [42,23]. For example, Isomap [31] finds a low-dimensional representation of the data by preserving geodesic distance between the samples. More

related to the techniques used in this work, local manifold learning techniques aim to maintain local structures in the data. Locally linear embedding (LLE) [25] represents local geometric attributes by reconstructing a sample linearly using its nearest neighbours. Laplacian eigenmaps [3] uses locally pair-wise Euclidean distances to characterise the local structure.

An important issue about a locally linear model is how to determine the accurate range of a local area in the data distribution. For a good prediction, the choice should achieve a balance between including rich information and taking high risk of mixing difference branches of the manifold. On the other hand, previous research of linear models has shown that constraints such as sparsity or non-negativity on the weights can help the model automatically choose proper predictor variables in training [33,15,12,32,16]. In particular, models regularised by sparsity constraint tend to be more robust and more interpretable [41]. Therefore, when constructing the locally linear prediction model, we impose sparsity constraints by employing the technique in [9].

To represent the textures in an image, previous works including our algorithms [21,20] consider a rectangle region at each pixel. To reduce the artefacts introduced by this arbitrary choice of the shape of a unit image region, we propose to include the semantic attributes of the pixels in the prediction model. The semantic attribute of a pixel is determined by the group to which the pixel is belongs. The groups are obtained by clustering orientation-invariant features [35]. The orientation-invariant features consist of eight maximum response (MR8) from a bank of 38 features. The responds of the filter bank reflects local contents in an image, and the maximisation provides the desired orientation-invariance. Intensive research has shown that clustering of representative local features of an image provides semantic analysis of the image. For example, global- and local-spatial features are used with fuzzy clustering for image segmentation [39]. Local features has been explored in recognising facial characteristics in [2]. Recently, models for document analysis has proved to be effective in dealing with visual data, i.e. images and videos [11], where feature groups serve as vocabulary of visual words and topic-based statistical models have been employed to deal with the data [5].

3. Model for colour prediction by examples

This section presents the proposed techniques for image colourisation. The section begins with a formal definition of the example-based image colourisation problem, which is followed by a discussion on the motivation of the proposed model. Then the prediction model is treated with the essential settings, i.e. manifold data distribution, sparsity and semantic grouping of pixels. We also consider necessary implementation details.

3.1. Colourisation problem

For the convenience of discussion, we first introduce some denotation conventions and frequently used symbols. We use latin letter "X/x" in symbols for variables related to gray-scale (luminance) information, and "Y/y" for those related to chromatic information, respectively. We attach a subscript "t" to symbols of variables related to the target image, i.e. the image for which we have its luminance information and want to infer the colours; correspondingly, we attach a subscript "s" to symbols of variables related to the source image, i.e. the reference example.

In particular, the task of colourisation is to predict the colours $\mathbf{Y}_t$ for a monochrome image $\mathbf{X}_t$, given a reference image $(\mathbf{X}_s,\mathbf{Y}_s)$. As specified above, $\mathbf{X}_s$ and $\mathbf{Y}_s$ represent the luminance (gray-scale intensities) and colour channels of the reference image, respectively; and $\mathbf{X}_t$ and $\mathbf{Y}_t$ represent those of the target image. The goal of learning is a representative model of the relations between the information in $\mathbf{X}_s$ and that in $\mathbf{Y}_s$. Thus for $\mathbf{X}_t$, the model produces estimated colours $\mathbf{Y}_t$.

As we have discussed in Section 1, it is generally not possible to find a unique pixel-wise luminance-colour mapping from an example colour image. On the other hand, the patterns of the luminance intensities in a small region provide clues of the content in the region. Given the example image, this information can help us to determine the colours in the region. In other words, at a particular position in an image, textual information eliminates the ambiguity about the potential colours and allows us to make a unique estimate. Therefore we consider the (overlapping) rectangular patches in an image and learn the relation between the texture and colour patterns at these patches. We let the symbols of the images denote the corresponding collection of patches and use lower-case symbols for individual patches, i.e. $\mathbf{X}_t \coloneqq \{\boldsymbol{x}_t^p\}_{p=1}^{N_t}$, $\mathbf{Y}_t \coloneqq \{\boldsymbol{y}_t^p\}_{p=1}^{N_t}$, $\mathbf{X}_s \coloneqq \{\boldsymbol{x}_t^q\}_{q=1}^{N_s}$ and $\mathbf{Y}_s \coloneqq \{\boldsymbol{y}_s^q\}_{q=1}^{N_s}$, where N_t and N_s are the number of patches in the grayscale image and the reference image, respectively.

A prediction model of the target colours can be formulated as

$$\boldsymbol{y}_t = f(\boldsymbol{x}_t \,|\, \{(\boldsymbol{y}_s^q,\boldsymbol{x}_s^q)\}_{q=1}^{N_s}), \tag{1}$$

where we have omitted the superscript $p=1,\ldots,N_t$ for the target patches. A shared linear model is arguably the simplest representation of the population of the texture and colour patterns. In particular, we assume the texture pattern $\boldsymbol{x}_t$ is linearly dependent on the reference texture patterns $\boldsymbol{x}_s$, and thus equals to a weighted sum $\boldsymbol{x}_t = \sum_q w_q \boldsymbol{x}_s^q$. Since we have assumed a shared geometric structure of the texture and the colour populations, the weights $\boldsymbol{w}=[w_1,\ldots,w_q]$ characterised the relation between the target patch $(\boldsymbol{x}_t,\boldsymbol{y}_t)$ and the reference patches $\{(\boldsymbol{y}_s^q,\boldsymbol{x}_s^q)\}$. Thus the estimation of the colours of the target patch is the linear combination of $\{\boldsymbol{y}_s\}$ by using weights $\boldsymbol{w}$. Eq. (1) is implemented by

$$\boldsymbol{y}_t = \sum_q w_q \boldsymbol{y}_s^q \tag{2}$$

$$\text{s.t.} \quad \boldsymbol{x}_t = \sum_q w_q \boldsymbol{x}_s^q,$$

which is a linear regression problem.

3.2. Manifold of patch populations

Despite the simplicity of the linear model of (2), it is obviously impractical for the patches in an image. Meaningful patches from real images consist of only a small fraction of all possible appearances of a region. Nevertheless, the population of real patches is *not* a linear subspace of the space of appearances. This is clear if we consider the fact that the superposition of two practical patches often produces meaningless result. Therefore, the globally linear structure in (2) is not a reliable representation of the texture and colour pattern populations, and the linear prediction model cannot give accurate estimation of the target colours.

On the other hand, rich research works have pointed out that the population of real image patches form a low-dimensional Riemannian manifold [8,36,30,6,10]. The variance of the appearances of the patches are controlled nonlinearly by a few latent semantic factors. It is sensible to assume that most semantic factors, e.g. distance to the camera sensor, spatial position of an object, occlusion relations between objects, etc., affect both luminance intensities (texture) and the colours in a similar manner. Thus our basic assumption that the texture and colour populations share a geometric structure remains viable if we adopt manifold distributions of the data. It is the assumption of a globally linear distribution that needs modification.

Fortunately, although presenting a non-linear structure globally, a Riemannian manifold has locally linear geometry. Each sample taken from a Riemannian manifold has a small neighbourhood,[1] which can be deemed as lying in a linear subspace. Therefore we limit the range of the linear relationship modelled by (2). More specifically, only a small number of the reconstruction weights $\{w_1,\ldots,w_{N_s}\}$ are allowed to be effective, i.e. nonzero. If there are K effective weights $\{w_{i_1},\ldots,w_{i_K}\}$, then the corresponding reference patches $\{i_1,\ldots,i_K\}$ are the K nearest neighbours to the target patch. In practical implementation, the neighbours are measured by the luminance,

$$\|\boldsymbol{x}_s^i-\boldsymbol{x}_t\| \le \|\boldsymbol{x}_s^j-\boldsymbol{x}_t\|$$

$$i\in\{i_1,\ldots,i_K\}, j\in\{1,\ldots,N_s\}\backslash\{i_1,\ldots,i_K\}.$$

Then we let the population of the colour pattern of the patches share the neighbourhood and the associated linear geometry, which are represented by $\{w_{i_1},\ldots,w_{i_K}\}$. The desired colours are estimated by

$$\boldsymbol{y}_t=\sum_{k\in\{1,\ldots,K\}} w_{i_k}\boldsymbol{y}_s^{i_k}, \tag{3}$$

where $\boldsymbol{y}_s^{i_k}, k=1,\ldots,K$ are the corresponding reference colour patterns. Note that the number of nearest neighbours is generally less than the dimension of the feature space of $\boldsymbol{x}_t$ and $\boldsymbol{x}_s$, thus the equation regarding the weights, $\boldsymbol{x}_t=\sum_q w_q\boldsymbol{x}_s^q$ becomes a least squared error minimisation

$$\boldsymbol{w}=\arg\min_{\boldsymbol{w}}\|\boldsymbol{x}_t-\sum_{k=1}^{K} w_{i_k}\boldsymbol{x}_s^{i_k}\|_2^2, \tag{4}$$

where $\|\cdot\|_p$ stands for p-norm. Note that the weight vector $\boldsymbol{w}$ in (2) and (4) are of different dimensions, N_s and K, respectively.

3.3. Sparsity constraints

To implement the locally linear prediction model (3), an important issue is to determine the size K of a neighbourhood. A big number of neighbours can improve the reliability of the reconstruction of the locally linear subspace, and can provide rich information of the colours. The downside of a large neighbourhood is that it increases the risk of intersecting different branches of a manifold, i.e. some samples in a oversized neighbourhood cannot be summarised in a linear structure. An oversized neighbourhood contains confusing colours and harms the prediction.

Choosing the number of nearest neighbours is a non-trivial task. First, without observing the data, we cannot have much confidence for any preset value to be close to the ideal size. Second, consider a better informed case, where we have some rough idea about a good neighbourhood size and choose K^* nearest neighbours at each patch. However, the ideal size of a neighbour often varies from one sample to another in practice. Third, given the overlapping nature of the patches in our problem, the nearest neighbours can be inappropriate choice of a neighbourhood. We consider an example to clarify this paradoxical statement. Let the neighbourhood size be determined to be some value, e.g. five, at some target luminance patch $\boldsymbol{x}_t$. Because we take overlapping patches from the reference image, it is possible that four out of the five nearest neighbours are patches that nearly duplicate each other and contribute essentially the same colour information. In this case, the colours $\boldsymbol{y}_t$ are estimated from two, instead of five, example patches, which can be insufficient.

Sparsity constraint has been shown effective for determining the complexity of a linear model automatically [33]. We adopt the sparsity constraint for our particular problem of determining a neighbourhood on an manifold to conduct linear regression. The neighbourhood size K is initially set to a relatively large value, and K nearest neighbours for each target patch are found accordingly. The tentative choice of the neighbourhood is trimmed by imposing sparse regularisation on the weights in the regression problem. In particular, we adopt the technique of least angle regression [9] to impose ℓ_1-penalty on the weights. The weights are obtained by minimising the ℓ_1-penalised least squared error problem (developed from (4))

$$\boldsymbol{w}=\arg\min_{\boldsymbol{w}}\left\|\boldsymbol{x}_t-\sum_{k=1}^{K} w_{i_k}\boldsymbol{x}_s^{i_k}\right\|_2^2+\gamma\|\boldsymbol{w}\|_1, \tag{5}$$

where the ℓ_1-penalty encourages the individual weights in $\boldsymbol{w}$ to shrink to 0, and consequently, the corresponding neighbours to cease affecting the prediction of colours. The parameter γ controls the strength of the sparsity penalty.

[1] Note that a neighbourhood in a patch manifold is not to be confused with the small surrounding rectangular area at a pixel in an image: the former consists of a subset patches in the population of all patches, and the latter makes one patch from an image.

3.4. Semantic groups

In the previous development, we have used rectangular patches of an image are the basic unit of learning and prediction. This practice is both convenient for implementation and consistent with convention in the literature. However, the rationale of the colourisation scheme is to infer colours based on the semantic information conveyed by the texture. Rectangular patches do not respect the boundaries between semantic contents in an image and may introduce artefacts in the prediction. In each patch, there can be multiple sub-regions belonging to different semantic categories. The shapes of those sub-regions vary from one patch to another, and are difficult to match between the target patch and the reference patches. Thus linearly combining the reference patches may mix colours from different semantic contents and result in artefacts.

Therefore we propose to use semantic information to assist the colour prediction. In particular, texton maps are computed for both the reference and the target image. In a texton map, a semantic category is assigned to each pixel by clustering features computed at the pixels. For features, the respondences to a bank of 38 filters are computed at each pixel: 18 edge filters (six orientations by three scales), 18 bar filters of the same settings, one Gaussian filter and one Laplacian filter. For the feature to represent semantic contents, orientation-invariance is a desired attribute. Therefore we employed the MR8 feature [35], where only the maximum respondences of the six orientations are taken at each scale for the edge and the bar filters. The texton map is obtained by clustering the orientation-invariance features by K-means.

Having computed the texton maps, we consider the consistence of the textons between a reference patch and the target patch when using the reference patch in prediction. The texton map in the reference patch are compared pixel-wisely with that in the target patch. Colours at the pixels with matched texton map are transferred from the reference patch by using the weights computed from the locally linear model. If all effective reference patches (with nonzero weights) are not matched by texton maps for a pixel, the weights of matched patches are re-normalised to compensate the mis-matched reference patches.

Let the texton map of the target patch $\boldsymbol{x}_t$ be $\boldsymbol{s}_t$, that of the reference luminance patch $\boldsymbol{x}_s^{i_k}$ be $\boldsymbol{s}_s^{i_k}$. Comparing $\boldsymbol{s}_t$ and $\boldsymbol{s}_s^{i_k}$, we have the matching map $\boldsymbol{u}^k \in \{0,1\}^{N_P}$ for the k-th reference patch, where N_P represents the number of pixels in a patch. The colour patch $\boldsymbol{y}_s^{i_k}$ is added to the prediction after being weighted by the corresponding weight w_k and being masked by $\boldsymbol{u}^k$.

3.5. Prediction algorithm

In summary, for an input grayscale patch $\boldsymbol{x}_t \in \mathbf{X}_t$, we can formulate the procedure of estimating the corresponding colours $\boldsymbol{y}_t$ as the following steps.

1. Compute orientation-invariant features for the target and reference image.
2. Cluster the features into textons. Obtain textion map for the target and reference image.
3. Find the K nearest neighbours of $\boldsymbol{x}_t$ in $\mathbf{X}_s$, $\{\boldsymbol{x}_s^{i_1},\ldots,\boldsymbol{x}_s^{i_K}\}$, where K is a relatively large number. The corresponding texton maps are $\boldsymbol{s}_s$ and $\{\boldsymbol{s}_t^{i_1},\ldots,\boldsymbol{s}_t^{i_K}\}$.
4. Compute K combination coefficients $\{w_{i_1},\ldots,w_{i_K}\}$ for $\{\boldsymbol{x}_s^{i_1},\ldots,\boldsymbol{x}_s^{i_K}\}$, respectively. The coefficients are obtained by solving an ℓ_1-penalised regression problem formulated in 5.
5. Synthesize $\boldsymbol{y}_t$ by combining the corresponding neighbours $\{\boldsymbol{y}_s^{i_1},\ldots,\boldsymbol{y}_s^{i_K}\}$ with the coefficients $\{w_{i_1},\ldots,w_{i_K}\}$ computed in Step 4 with matching mask as specified in Subsection 3.4.

Steps 3–5 are performed for each patch in the target image. In Step 5, the matching mask is a binary array of comparing the texton maps of the target and the reference patches at each pixel,

$$\boldsymbol{u}^k(l) = \begin{cases} 1, & \boldsymbol{s}_s^{i_k}(l) = \boldsymbol{s}_t(l) \\ \epsilon, & \text{otherwise,} \end{cases}$$

where $\epsilon \approx 0$ is a tiny positive number for normalisation and $l = 1,\ldots,N_P$ the index for the pixels in a patch. The prediction is completed by

$$\boldsymbol{y}_t = \sum_{k\in\{1,\ldots,K\}} \tilde{\boldsymbol{w}}_{i_k} \odot \boldsymbol{y}_s^{i_k}, \tag{6}$$

where $\odot$ represents element-wise product and

$$\tilde{\boldsymbol{w}}_{i_k}(l) \coloneqq \frac{\sum_k w_{i_k}\boldsymbol{u}^k(l)}{\sum_k \boldsymbol{u}^k(l)},$$

for $l = 1,\ldots,N_P$. The pixel-wise weights $\tilde{\boldsymbol{w}}_{i_k}$ combines the reconstruction weights and texton matching. The small conditioning number ϵ is introduced for singular pixels, where the texton at a singular pixel does not find a match in any effective reference patch with a non-zero weight. For the singular pixels, the colours are the average of the colours at the corresponding pixels in all effective reference patches.

4. Experiment

This section presents some further technical details about the implementation of the proposed scheme and demonstrates experimental results showing effectiveness of our algorithm. We show examples of how the adopted settings help us to improve the colourisation, compared to previous methods. We also report subjective study of the results. All figures in our report are better viewed in colours on a computer screen.

4.1. Implementation details

Feature vectors of patches: The luminance information of a patch is represented by feature vector $\boldsymbol{x}$, which is constructed as in [6,7]. The feature vector $\boldsymbol{x}$ consists of three components: the average pixel intensity, the first and the second order intensity gradients at individual pixels. This composition reflects both the overall luminance and textures in the patch.

Formally, the feature vector of an image patch is built as follows. An image can be considered as a function $\mathcal{I}:\mathbb{Z}^2\to\mathbb{R}$. The horizontal and vertical differentiation operators are defined as

$$\nabla_x\mathcal{I}(x,y)=\mathcal{I}(x+1,y)-\mathcal{I}(x-1,y)$$

$$\nabla_y\mathcal{I}(x,y)=\mathcal{I}(x,y+1)-\mathcal{I}(x,y-1). \quad (7)$$

The feature vector of a grayscale patch $\mathcal{P}$ is then defined as

$$[\lambda\overline{\mathcal{I}|_{\mathcal{P}}}\ \ \nabla_x\mathcal{I}|_{\mathcal{P}}\ \ \nabla_y\mathcal{I}|_{\mathcal{P}}\ \ \nabla_x^2\mathcal{I}|_{\mathcal{P}}\ \ \nabla_y^2\mathcal{I}|_{\mathcal{P}}]^T,$$

where the first element represent the average pixel intensity in that patch

$$\overline{\mathcal{I}|_{\mathcal{P}}}=\frac{\sum_{(x,y)\in\mathcal{P}}\mathcal{I}(x,y)}{|\mathcal{P}|},$$

and λ is the weight of the intensity. The weight should be chosen according to the patch size, which is to keep the balance between the influence of the average intensity (representing luminance) and the gradients (representing the texture details). The weight is necessary, because the entries of the gradients in the feature vector grows with respect to the size of the patch, but the average intensity is a scalar irrelevant to the size of a patch. The colour patches are the hue/saturation values at individual pixels in the corresponding patches. In our implementation, λ is set to the number of pixels in a patch N_P, which gives satisfactory results (Fig. 1).

Fig. 2 demonstrates the construction of the feature vector for an input grayscale patch and that for its three nearest neighbours in the reference grayscale patches.

Fig. 2(a) and (b) shows the input and the reference images, respectively. The blueish box in (a) indicates an input patch, whose colours are to be predicted. The first column in Fig. 2(c–g) represents the feature of the query patch: the intensity values, the horizontal gradients, the vertical gradients, the horizontal second order gradients and the vertical second order gradients, respectively.

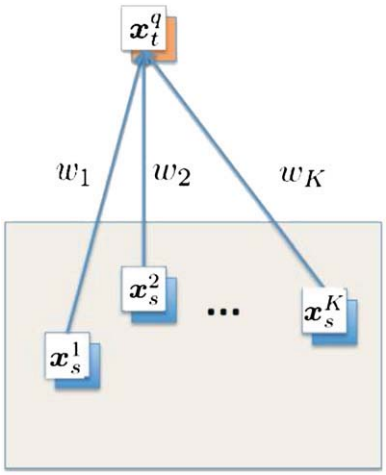

Fig. 1. Flow chart of locally linear embedding with sparsity constraint. The white boxes indicates grayscale patches. The big enclosing gray box indicates the reference image. The blue boxes behind each grayscale patch represents the colour information of those patches. After the coefficients $w_1,\ldots,w_K$ have been computed, they are used to recover the colours of the query patch (shown in the orange box). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

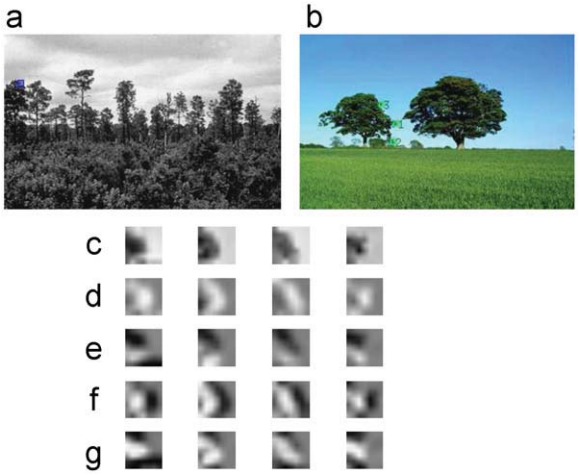

Fig. 2. Nearest neighbours. (a) Input; (b) training; (c)–(g) features. (Courtesy of the authors of [21]).

Three patches with closest features to the query feature has been shown in Fig. 2(b) by greenish boxes. Their feature components are shown in the second, the third and the last column in Fig. 2(c–g).

Parameter and algorithm settings: In practical implementation of the colourisation, there are several parameters to be determined. We discuss the settings of these parameters for completeness of our introduction.

The first implementation setting is the size of the patches. Large patches contains rich textural information. But large patches tend to be more specialised for a particular image; and it is more difficult to find good correspondence in a reference image. On contrary, if the patch size is too small, the textures in many patches are ambiguous; and it is difficult to choose relevant neighbours from many "good" correspondences. We resize our images so that the edges are of 300 to 400 pixels. Then we have found the patch size between 5×5 to 15×15 gives satisfactory results. We use the patch size of 5×5 in our experiments. In our tests, the overlapping is set to two pixels less than the patch size.

For the tentative neighbourhood size K, we have found that 20 to 100 is a reasonable range in our experiments.[2] The sparsity constraint is configured so that $K/4$ neighbouring patches are effectively contributing to the prediction.

Finally, the number of textons, i.e. that of the clusters of pixel-wise orientation-invariant features, affects the fineness of semantic partition we consider when constructing the colours. Our algorithm is stable over a range of choices, we have used 8 in our experiments.

Fig. 3 demonstrates the procedure of choosing reference patches for a particular configuration of the algorithm. The input grayscale image and the reference example are the same as we have shown in Fig. 2. In

[2] The total number of reference patches is about 10,000 to 15,000 for the chosen image and patch sizes and overlap.

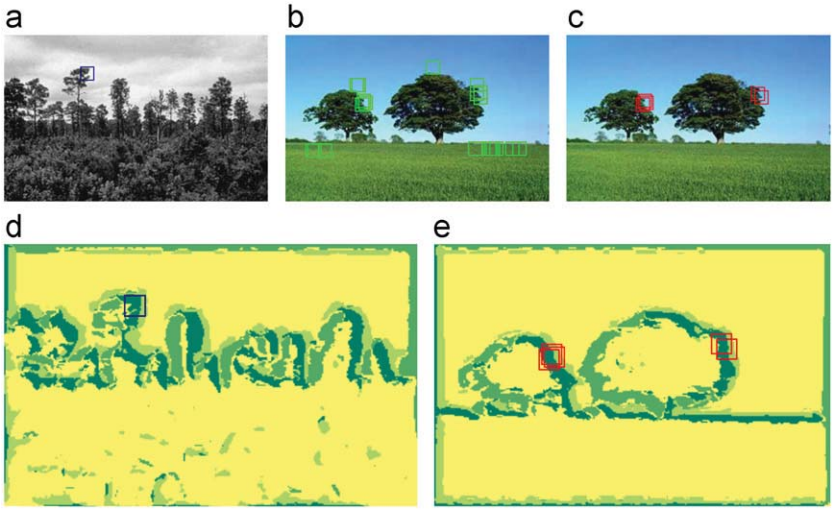

Fig. 3. Choose reference patches. (a) Input image, blue box indicates a patch of interest; (b) reference image and 20 nearest reference patches to the target patch in (a), marked by green boxes; (c): effective (with non-zero weights) reference patches, marked by red boxes; (d): texton map of the input image, target patch is marked as in (a); (e): texton map of the reference image, target patch is marked as in (c). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

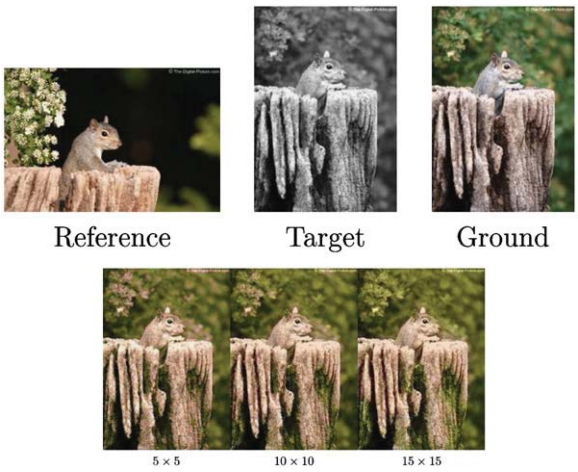

Fig. 4. Colourisation with different patch sizes. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

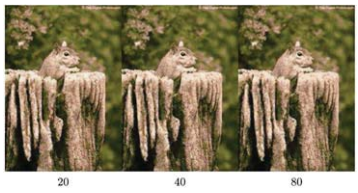

Fig. 5. Colourisation with neighbourhood sizes. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

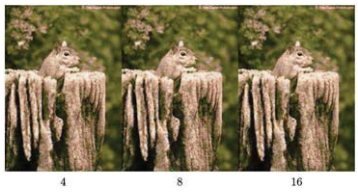

Fig. 6. Colourisation with different number of textons. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Fig. 3, we use patches of 15×15 to make the illustration clear. Fig. 2(a) shows the input image with a patch of interest. Fig. 2(b) shows 20 nearest neighbours found for the target patch in the reference image. Fig. 2(c) marks only the effective reference patches, i.e. patches with non-zero weights. In (d) and (e) we give the texton maps of the input and the reference image with marked relative patches. The figure indicates that the sparsity constraint is effective in selecting useful reference patches. Moreover, even if an irrelevant patch passed the sparsity checking (consider a patch of the grass in (b) passed through into stage (c)), the texton map of the patch is unlikely matched with the target patch, and most of the pixels in the irrelevant reference patch will be masked off.

Figs. 4–6 show some colourisation results with different configurations of the algorithm. The target and the reference images are shown in Fig. 4, as well as the ground truth of the colours in the target image. Those results demonstrates that the algorithm performed stably in the tested settings. This is consistent with our discussion above.

4.2. Experiment results

We use two examples to show how the proposed treatments improve the colourisation performance. Fig. 7 displays the predicted colours for the input and the reference images in Fig. 4.

The image in (a) is resulted by the method of [38], where the colours of the subject are not distinguishable from those of the background. Fig. 7(b) shows that the

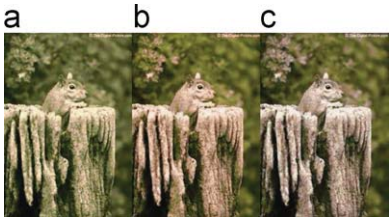

Fig. 7. Transfer colours to an image of a squirrel.(For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

introduce of texture information and manifold assumption of the data distribution improves the colourisation [21]. Fig. 7(c) shows the colours recovered by the sparse model [20]. Both results (b) and (c) are satisfactory, as they apply correct colour tunes to the squirrel and the background respectively. However, result (c) demonstrates sharper contrast in colour tunes between the subject and the scenery. A possible explanation is the contribution from the sparsity constraint. For each patch in the input image, the sparse constraint limits the number of patches in the training image that can transfer their colours to the input patch. Therefore the constraint reduces the mix of colours and produce distinctive colours for different components in the image. This is consistent with the example of selecting effective reference patches as we have seen in Fig. 3.

The next example shows the usefulness of the introduction of texton maps. A second example is needed to emphasise the contribution of texton maps, because the benefit from texton maps is similar to and complementary for that from the sparsity constraint: both sparsity constraint and texton maps help reduce mixing of irrelevant colours in the prediction. Fig. 8 demonstrates the effect of the texton maps for recovering the colours of a tree. The texton maps worked as expected as the discussion above (c.f. Fig. 3) and helped us to eliminate the artefacts indicated by the red arrows. We can see that when the neighbourhood size is big, texton maps can assist the sparsity constraint to eliminate irrelevant colours.

Fig. 9 shows more examples of the colourisation results. The effectiveness of the proposed method can be observed from most of those results. It is worth noting that the predicted colours may not be faithful replica of the ground-truth colours, but reflect the colour tunes in the reference image. However, since the reference image is a natural image containing the similar content as the that in the target image, the predicted colours remain satisfactory to an observer's eyes. This effect is clear in the example of the swan images.

As an objective evaluation, we also test the proposed techniques for predicting *missing-colours* in images. That is we set a region of an image into monochromatic, then use the remaining colours of the image to predict the removed colours. This testing scheme can avoid irrelevant variations in a practical scenario, and thus is useful for assessing the merits of the techniques. In particular, we take a region of $[40 \times 80]$ from images of $[180 \times 270]$ for

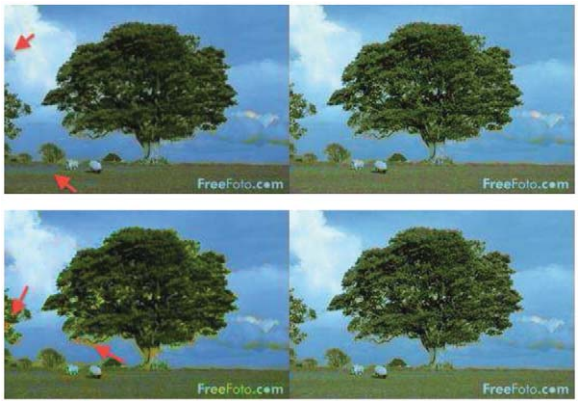

Fig. 8. Transfer colours w/wo texton maps. Top: neighbourhood size is 20; Bottom: neighbourhood size is 80. The effect of texton maps becomes more obvious when the neighbourhood size is big. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

the test. The test is performed on 10 image classes consisting of totally 100 images. The predicted and the true colours are compared and the squared errors of the saturation and hue channel are recorded. Four schemes are tested: (i) the technique of [38], (ii) the manifold-based regression scheme in (3) and (4), (iii) sparse (5) based on (ii), and (iv) semantic texton map introduced based on (iii). The errors on each image classes are normalised against the mean error of scheme (i) for a clear display. Fig. 10 shows the progressive improvement on the performance introduced by the proposed techniques. The overall performance is shown in the figure as well.

It is generally difficult to develop an objective criterion to assess the result. Even if the ground-truth colours of the target image is available, an algorithm cannot be simply judged by comparing the predicted colours with the ground-truth colours. This is because the result of colourisation is affected by both the textures in the target image and the colours in the reference image, and the lighting conditions may differ from the reference and the target images. Furthermore, from a wider perspective, colourisation belongs to the family of techniques of visualisation. The result should be judged by the end users, who are human observers. Therefore, we employ a group of volunteers to assess the colourisation results given by the proposed method and that by the previous method [38]. We use both methods to colourise 100 images, for which some examples are shown in our earlier discussion. Then for each image, we let a subject select the preferred result or indicate that there is no obvious preference for the image. Figs. 11 and 12 summarised the results of the subjective study. Fig. 11 displays the preference by each subjects. Fig. 12 is a histogram of the preference of the images. For each image, the preference is determined by the number of subjects who preferred the proposed method subtracted by the number of subjects who preferred that by [38]. Thus an image can have a score in $[-\#.\text{subjects}, +\#.\text{subjects}]$, higher score indicates stronger preference of the proposed method.

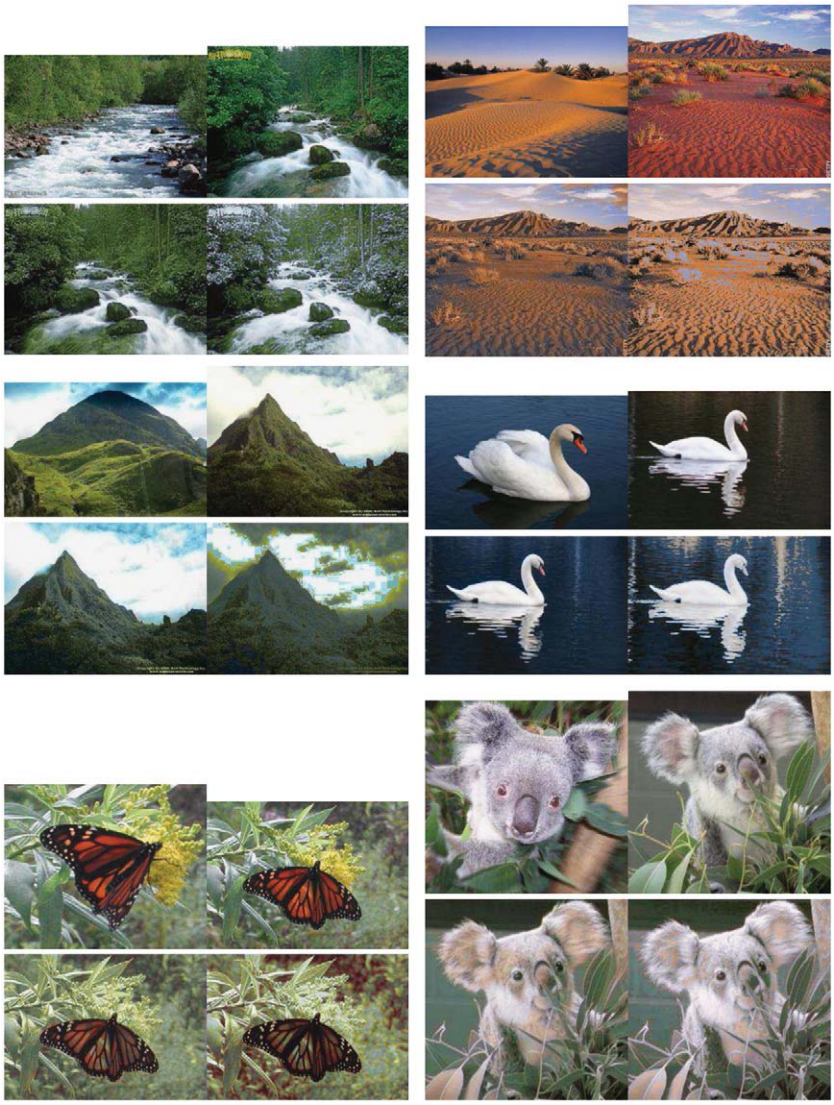

Fig. 9. Prediction colours. In each sub-figure, the top-left pane shows the reference image, the top-right pane shows the grand-truth target image, the bottom-left pane displays the colourisation result by the proposed method, and the bottom-right pane compares the colours resulted by the method in [38]. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

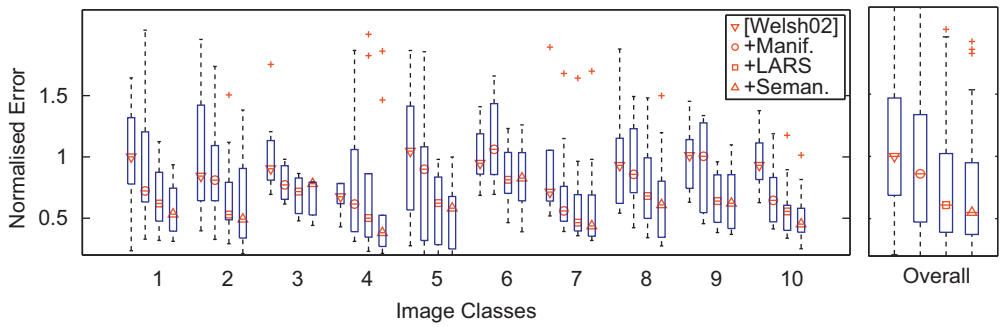

Fig. 10. Prediction of missing colours. **[Welsh02]**: scheme (i); **Manif.**: scheme (ii) based on manifold assumption; **LARS**: scheme (iii), introducing sparsity; **Seman.**: scheme (iv), introducing semantic texton map. See text for detailed discussion.

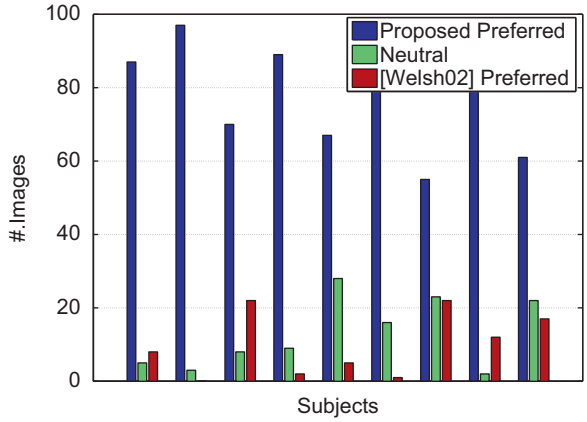

Fig. 11. Subjects' preference of predicted colours.(For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

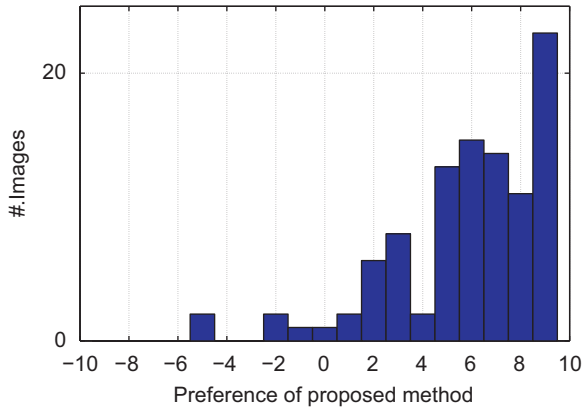

Fig. 12. Histogram of preference of predicted colour images.

The two figures clearly demonstrated the improvement achieved by the proposed colourisation technique.

5. Conclusion

In this paper, we have proposed an automatic method of learning the relations between colours and texture in an image, and transferring colours to a grayscale image by exploiting the learned relations. The method is based on the manifold assumption about the geometric property of the texture distributions. We also employ sparsity constraint to determine the interested local area of the manifold. The prediction is assisted by semantic maps of the image, which are obtained via clustering robust features.

At the theoretical level, three branches of previous research precedes the present work: the study about the geometry of the image feature space [4,10], the study of sparse algorithms [33,9], and the study of invariant image features and the associated segmentation techniques [35]. The contribution of the current work is at the methodological level. We integrate those methods as model analysing the relations between colours and textures in images. The model is then applied to the prediction of colours for grayscale images using examples; and useful results are obtained.

Compared to existing research on the problem of colourisation of grayscale images, the proposed method has several advantages. It involves small amount of human effort, and remains stable in a reasonable range of the configurations of the algorithm. We have also employed human users for a subjective study. The reported objective evaluation and user experience has supported the proposed method. Future research will focus on a unified formulation of the model, preferably within a probabilistic framework, which provides principal justification of the system from the statistic perspective.

References

[1] M. Abbasnejad, D. Ramachandram, R. Mandava, A survey of the state of the art in learning the kernels, Knowledge and Information Systems 31 (2) (2012) 193–221.
[2] M.E. Aroussi, M.E. Hassouni, S. Ghouzali, M. Rziza, D. Aboutajdine, Local appearance based face recognition method using block based steerable pyramid transform, Signal Processing 91 (1) (2011) 38–50.
[3] M. Belkin, P. Niyogi, Semi-supervised learning on riemannian manifolds, Machine Learning 56 (2004) 209–239.
[4] D. Beymer, T. Poggio, Image representation for visual learning, Science 272 (1996) 1905–1909.
[5] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
[6] H. Chang, D. Yeung, Y. Xiong, Super-resolution through neighbor embedding, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004.
[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.
[8] D.L. Donoho, C. Grimes, Image manifolds which are isometric to euclidean space, Journal of Mathematical Imaging and Vision 23 (1) (2005) 5–24.
[9] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Annals of Statistics 32 (2) (2004) 407–451.
[10] W. Fan, D.-Y. Yeung, Image hallucination using neighbor embedding over visual primitive manifolds, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.
[11] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.
[12] N. A. O. G. M. for Nonnegative Matrix Factorization, Online nonnegative matrix factorization with robust stochastic approximation. IEEE Transactions on Signal Processing 60 (6) (2012) 2882–2898.

[13] B. Geng, D. Tao, C. Xu, Daml: domain adaptation metric learning, IEEE Transactions on Image Processing 20 (10) (2011) 2980–2989.
[14] B. Geng, D. Tao, C. Xu, L. Yang, X.-S. Hua, Ensemble manifold regularization, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (6) (2012) 1227–1233.
[15] N. Guan, D. Tao, Z. Luo, B. Yuan, Online nonnegative matrix factorization with robust stochastic approximation, IEEE Transactions on Neural Networks and Learning Systems 23 (7) (2012) 1087–1099.
[16] K. Huang, Y. Ying, C. Campbell, Generalized sparse metric learning with relative comparisons, Knowledge and Information Systems 28 (1) (2011) 25–45.
[17] X. Jing, S. Li, C. Lan, D. Zhang, J. Yang, Q. Liu, Color image canonical correlation analysis for face feature extraction and recognition, Signal Processing 91 (8) (2011) 2132–2140.
[18] C. Keßler, What is the difference? a cognitive dissimilarity measure for information retrieval result sets, Knowledge and Information Systems 30 (2) (2012) 319–340.
[19] Y. Kim, C.-W. Chung, S.-L. Lee, D.-H. Kim, Distance approximation techniques to reduce the dimensionality for multimedia databases, Knowledge and Information Systems 28 (1) (2011) 227–248.
[20] J. Li, W. Bian, D. Tao, C. Zhang, Learning colours from textures by sparse manifold embedding, in: Australasian Joint Conference on Artificial Intelligence, 2011.
[21] J. Li, P. Hao, Transferring colours to grayscale images by locally linear embedding, in: Proceedings of British Machine Vision Conference, 2008.
[22] Z. Lin, M. Lyu, I. King, Matchsim: a novel similarity measure based on maximum neighborhood matching, Knowledge and Information Systems 32 (1) (2012) 141–166.
[23] J. Lu, Y. Zhao, Dominant singular value decomposition representation for face recognition, Signal Processing 90 (6) (2010) 2087–2093.
[24] Y. Qu, T.-T. Wong, P.-A. Heng, Manga colorization, in: Proceedings of Siggraph, 2006.
[25] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[26] H.S. Seung, D.D. Lee, The manifold way of perception, Science (2000) 2268–2269.
[27] J. Silberg, Cinesite press article (1998). 〈 http://www.cinesite.com/core/press/articles/1998/10_00_98-team.html 〉.
[28] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, C. Zhang, Probabilistic exposure fusion, IEEE Transactions on Image Processing 21 (1) (2012) 341–357.
[29] M. Song, D. Tao, C. Chen, X. Li, C.W. Chen, Color to gray: visual cue preservation, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1537–1552.
[30] R. Souvenir, Manifold Learning for Natural Image Sets, Ph.D. Thesis, 2006.
[31] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[32] X. Tian, D. Tao, Y. Rui, Sparse transfer learning for interactive video search reranking, ACM TOMCCAP.
[33] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (1996) 267–288.
[34] L.J.P. van der Maaten, E.O. Postma, H.J. van den Herik, Dimensionality Reduction: A Comparative Review, Technical Report, Tilburg University, 2009.
[35] M. Varma, A. Zisserman, A statistical approach to texture classification from single images, International Journal on Computer Vision 62 (2005) 61–81.
[36] J. Verbeek, Learning non-linear image manifolds by combining local linear models, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (8) (2006) 1236–1250.
[37] F. Wang, P. Li, A. König, M. Wan, Improving clustering by learning a bi-stochastic data similarity matrix, Knowledge and Information Systems 32 (2) (2012) 351–382.
[38] T. Welsh, M. Ashikhmin, K. Mueller, Transferring color to greyscale images, in: Proceedings of Siggraph, 2002.
[39] F. Zhao, L. Jiao, H. Liu, X. Gao, A novel fuzzy clustering algorithm with non local adaptive spatial constraint for image segmentation, Signal Processing 91 (4) (2011) 988–999.
[40] Q. Zhu, E. Keogh, Mining historical manuscripts with local color patches, Knowledge and Information Systems 30 (3) (2012) 637–665.
[41] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, Journal of Computational and Graphical Statistics 15 (2) (2004) 262–286.
[42] W. Zuo, H. Zhang, D. Zhang, K. Wang, Post-processed LDA for face and palmprint recognition: What is the rationale, Signal Processing 90 (8) (2010) 2344–2352.