IET Computer Vision

Special Issue on Unsupervised Feature Learning for Vision

IET Journals
The Institution of Engineering and Technology

Strategy for aesthetic photography recommendation via collaborative composition model

ISSN 1751-9632
Received on 25th August 2014
Revised on 22nd January 2015
Accepted on 19th February 2015
doi: 10.1049/iet-cvi.2014.0276
www.ietdl.org

Yanhao Zhang[1], Qingming Huang[1,2] ✉, Lei Qin[3], Sicheng Zhao[1], Xiusheng Lu[1], Xiaoshuai Sun[1], Hongxun Yao[1]

[1]*School of Computer Science, Harbin Institute of Technology, Harbin 150001, People's Republic of China*
[2]*School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China*
[3]*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, People's Republic of China*
✉ *E-mail: qmhuang@jdl.ac.cn*

Abstract: In this study, the authors propose a collaborative composition model for automatically recommending suitable positions and poses in the scene of photography taken by amateurs. By analysing aesthetic-aware features, the authors' strategy jointly takes attention and geometry composition into account to learn the aesthetic manifestation knowledge of professional photographers. Firstly, aesthetic composition representation exploits the strength of visual saliency to explicitly encode the spatial correlation of the professional photos. Secondly, $\ell 2$ regularised least square is adopted to constrain the representation coefficients, which provides a fast solution in selecting aesthetic candidates collaboratively. In addition, a novel confidence measure scheme is further designed based on reconstruction errors and the reference photos are updated adaptively according to the composition rules. Both qualitative and quantitative evaluations show that the model performs well for the portrait photographing recommendation.

1 Introduction

With the increasing popularity of digital devices, there are massive numbers of photos shared in social websites, such as myphoto (http://www.photo.net/) and DPchallenge (http://www.dpchallenge.com/). Compared to traditional multimedia techniques, such data also feature in their aesthetic context, such as photographing locations and photographing pose, facilitating many photography-based services, such as aesthetic photograph assessment and photographic quality enhancement. In daily life, most portrait photos are taken by the users following their casual feelings in terms of usual photographing. For most amateurs, when taking photos in a great landscape or before a landmark, they always ask themselves how to pose or where to stand in the beautiful scenes. In terms of behaviour analysis and amateur assistance, such trends further emerged with the requirement of aesthetic photography, enabling the potential application to exploit expert knowledge for intelligent portrait photography. From a professional point of view, composition rules contribute a lot to the quality of the photos, considering the arrangement of the spatial context [1]. Experienced photographers adhere the aesthetic elements and rules to the artworks, which makes the photos more visually appealing. Professional photos are therefore taken using professional photography techniques and consistent with aesthetic composition rules. If the users could get such recommendations from similar professional works, it will fully satisfy users' needs during their amateur photography. Moreover, knowledge of professional photos could help the users to shoot better pictures and enrich the users' photography experience.

Motivated by this situation, we propose a novel intelligent photography approach to recommend the most suitable poses and positions. As illustrated in Fig. 1, we try to give an instant suggestion before the amateur photographer actually captures the photo by performing a strategy to recommend the pose and position in the given live preview. Such recommended photos can benefit from techniques of professional photographers, such as the rich photography experience of composition and arts. This goal is challenging because of multiple factors, such as the lack of training samples, the difficulty in aesthetic composition modelling, as well as the open problem in representing the given scene using reference photos. To this end, we resort to the knowledge from social multimedia by using professional photos crawled from photo-sharing websites. Specifically, we introduce the main difficulties that we are dedicated to solve as the following two problems: (i) How to model the aesthetic composition in the images considering that professional photos also suffer pose variance, zooming and viewpoint changes. That is, the reference and query images should be effectively represented regarding the advantages of the composition features. (ii) How to efficiently and accurately recommend pose and position given a query photo according to the reference data. That is, the pose and position for the query photos should be obtained from the recommended candidate results.

The aim of our strategy is to automatically recommend human poses and positions for amateur photography, borrowing the aesthetic professional knowledge. More specifically, an aesthetic composition representation is presented based on the attention and the geometry distribution of professional photos, which act as the reference for aesthetic composition components regarding the visual saliency and the partition preference. Followed by an effective solution towards robust selection of aesthetic candidates, we manage the aesthetic composition features by clustering analysis of reference photos. Subsequently, we propose a collaborative composition model based on $\ell 2$ regularisation to represent the query images given the clusters of the professional candidates. For recommendation, the live preview is sent as a query to select the clustered candidates, by which the most suitable photos are recommended instantly using reconstruction error measurement, making suggestions based on the composition rules.

Our contributions are threefold: (i) we propose a new perspective for photography recommendation, which is the first work that touches the point of aesthetic composition to recommend photography poses and positions; (ii) the collaborative composition model learnt from reference professional photos is leveraged, which not only exploits the strength of subspace representation, but also explicitly takes the attention and geometry into consideration and (iii) the proposed model makes use of reconstruction errors to

© The Institution of Engineering and Technology 2015

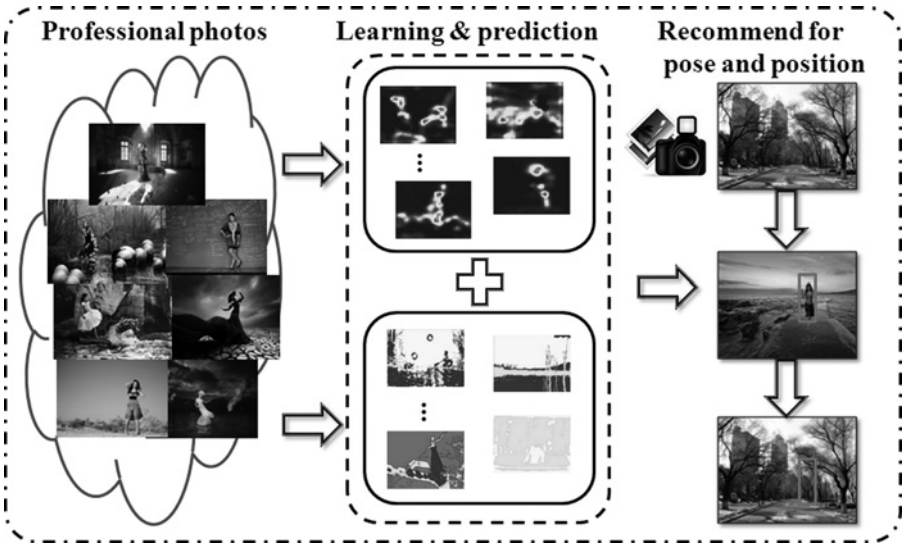

Fig. 1 *Overview of our proposed recommendation strategy*
We learn the aesthetic knowledge from professional photos and predict suitable poses and positions by the proposed collaborative composition model

effectively select reference candidates for a query image. A reasonable update mechanism that implicitly considers the rules of composition further helps to determine a suitable recommendation.

In the following, we review the related work on aesthetic-based assessment and image quality analysis in Section 2. Before illustrating the technical details, we show the overview framework in Section 3. Aesthetic composition representation is introduced in Section 4. Then, Section 5 presents the recommendation strategy based on our collaborative composition model. We give experimental evaluations in Section 6. Finally, we conclude this paper.

2 Related work

Aesthetic photography and photo-quality assessment are popular and extremely challenging topics. Previous approaches dealing with the content-aware photo-management problems mainly evaluate the quality of photos according to the photographic principles. Many interesting applications including perceptual photo-quality assessment, photo view recommendation and quality-based photo re-ranking have drawn great attention in the multimedia research community. Li *et al.* [2] proposed a framework of automatically evaluating the aesthetic quality on the photos with faces. For aesthetic appeal assessment in [3], low-level image composition features are used to build the aesthetic classifier. Some works have attempted to make personalised image tag recommendation [4] and photo selection towards aesthetic [5]. On the basis of aesthetic-related feature analysis, the above applications show great potential to improve the perceptual photo quality and guide the users to take photos that follow the aesthetic rules.

In terms of image quality analysis and image aesthetics analysis, the existing works concern image degradation caused by noise, distortion, and compression artefacts. Previous works on image quality assessment [6, 7] usually required the original undistorted image to assess the quality of the degraded image that caused by noise. Subsequently, some works [8, 9] have been proposed for directly estimating the quality of a single image. The image aesthetics analysis method [10] tried to design optimised visual features to mimic human perception and aesthetics criteria for the photo assessment. Ke *et al.* [11] designed a high-level semantic feature set for describing the spatial distribution of edges, blur and colour to model the people's perception of photo quality. In [12, 13], visual features are utilised with standard learning algorithms to model people's subjective evaluation for images. Luo and Tang [14] extracted and scored subject regions using a number of high-level semantic features to assess aesthetics of the photos. Sun *et al.* [15] detected the salience region of a photo, and analysed the aesthetics of the photo with the relative position of subject regions and other aesthetic properties. Bhattacharya *et al.* [16] proposed a framework for photo quality assessment and enhancement based on visual aesthetics. With a learning-based support vector regression model, the proposed system helps users to recompose the photos to improve its aesthetic quality. Xu *et al.* [17] studied the problem of mobile photography recommendation and proposed a learning-based method from massive social images.

Different from the previous works considering the aesthetic features of the entire image, our strategy focuses on aesthetic composition representation by emphasising the attention and geometry composition of professional photos. Similar to [1], the overview of the strategy we propose (shown in Fig. 1) can equip the digital cameras with smart function to help the users shoot near professional photos. The above scenario is essentially different from the existing works in making use of the user-sharing multimedia on the web. In terms of enhancing user photographing qualities, previous methods are more focusing on quality enhancement, while some of them try to tackle the problem of quality evaluation [14, 15]. Our scenario differs from recent works in terms of photography suggestion, which are trying to guide people to capture aesthetically pleasing photos [1]. Cheng *et al.* [1] proposed an intelligent photography system by analysing the locations and context of the objects. By learning the composition rules of professional photos, photos can be automatically and professionally generated or recommended from a wide or continuous view. Rather than suggesting the users to take a totally different image [18], we aim to instantly guide the photography by suggesting suitable positions and poses in a given scene. To evaluate our model quantitatively, we have conducted a group of experiments on the dataset that collected over 200 photos from social websites. The proposed model significantly outperforms the baseline methods. Besides, such applications will be exciting for digital camera users by giving automatic recommendation of 'How should I pose' when taking photos.

3 Method overview

Fig. 2 gives a pipeline of the proposed framework. To accomplish the goal of aesthetic photography, crawled professional photos from online sharing websites are first collected to form a reference photo dataset. Then, two components are exploited to obtain the aesthetic representation from professional photos based on attention and geometry composition, respectively. When a query

IET Comput. Vis., 2015, Vol. 9, Iss. 5, pp. 691–698
© The Institution of Engineering and Technology 2015

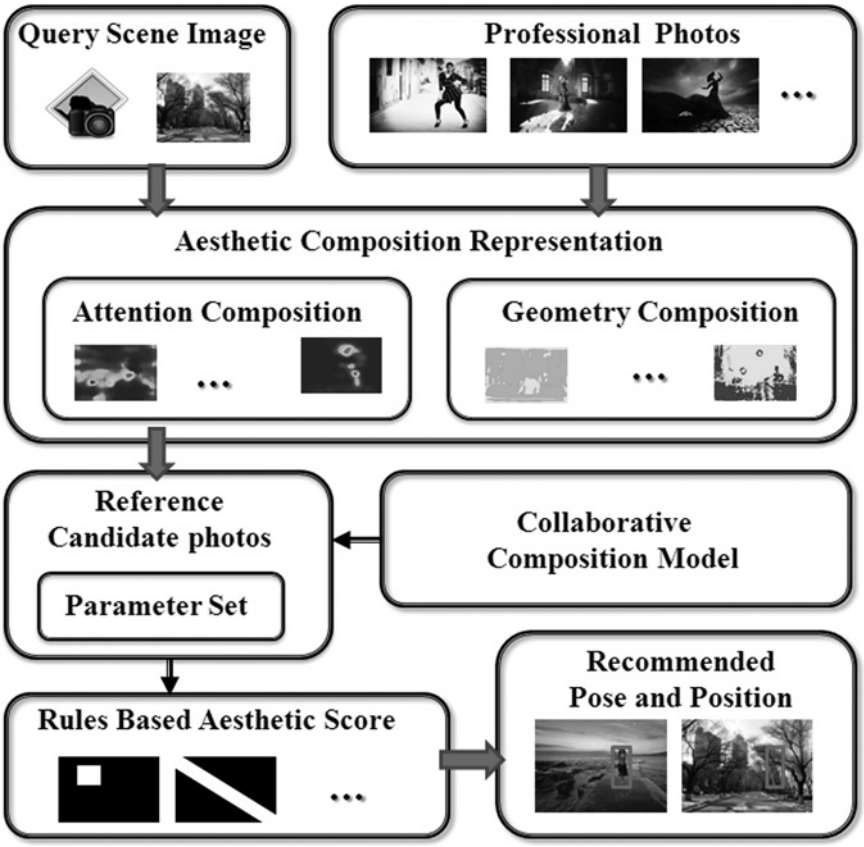

Fig. 2 *Pipeline of proposed recommendation strategy*

Main parts are focusing on aesthetic composition representation and the strategy of recommendation using collaborative composition model

photo without portrait is presented, we retrieve the reference photos in the database using a collaborative composition model. Finally, we select a single reference photo from the candidates adaptively with suitable poses and positions. Therefore, the recommendation is produced from professional photographs and consistent with classical composition rules, so is able to guide the users to take satisfactory photos.

4 Aesthetic composition representation

In this section, we discuss two composition features to obtain the aesthetic composition representation. The goal is to select the professional photos that share the most similarities with the query image on aesthetic composition. At the beginning, we use the human pose estimation method for still images [19] to estimate poses automatically in each professional photo. We re-annotate poses and positions of the photos with bounding box. Fig. 3 shows some examples from the dataset with ground truth positions and poses. The annotations associated with each professional photo are presented for recommendation.

The details of aesthetic composition representation are described below, consisting of attention and geometry composition features. The results of these two features cover most of the aesthetic expressions of professional photos, such that the poses and positions are worth consulting.

4.1 Attention composition feature

Visual saliency-based techniques are adopted to obtain the suitable estimation of spatial distribution in the subject regions. As we know, saliency maps with or without main subjects will be quite different. However, the main subject in the saliency map is also a component of the spatial attention composition. In order to represent the scene structure without stressing the portrait, we design a decaying exponential function to weaken the magnitude of saliency but preserving spatial attention distribution by (1). We use a PQFT method [20] to obtain initial saliency map S, expressing the salient region distribution

$$S^{\mathrm{d}} = S \times f(S(x, y)) \tag{1}$$

where S^{d} refers to the decay saliency map, $\overline{S}$ is the average value of S. $f(S(x, y))$ is the indicator of the decay saliency

$$f(S(x, y)) = \begin{cases} 1 & S(x, y) < \overline{S} \\ \exp\left(-\boldsymbol{\beta} S(x, y)\right) & S(x, y) \geq \overline{S} \end{cases} \tag{2}$$

When $S(x, y)$ is greater than $\overline{S}$, we weaken the highest magnitude to make the uniform distribution. In this way, the decay saliency map can be used to represent the attention composition of the professional photos. For simplicity, we use decay saliency maps as the feature to represent the attention distribution of reference images where each image is normalised to 32×32 pixels. To make it possible to match in the feature space, we quantise the decay salient map of each photo S^{d} into a 1024-dim vector. In order to maintain all the spatial information, we resize the photo to scale $S_i'(x, y)$. For each image I_i in professional photo set Γ, a 1024-dim vector $\boldsymbol{U}_i$ is formed by concatenating the value of $S_i'(x, y)$.

4.2 Geometry composition feature

Besides the attention, geometry composition is also an important cue to measure the scene structure composition. It provides an additional discriminative amendment for aesthetic composition representation. For the professional photos set Γ, we attempt to use hill-climbing algorithm for colour image segmentation [21]. All of the photos are assumed to have less than 10 regions of

© The Institution of Engineering and Technology 2015

Fig. 3 *We annotated the poses and positions of professional photos as the labelled ground truth*
Main parts of the body, such as arms and legs are marked in the squares of different colours

segmentation. So we set the total region number as 10 for normalisation. For each region, we use $C = (\text{norm}(x), \text{norm}(y), \text{norm}(\text{area}(x, y)))$ as its 3-dim compositional feature, where area(x, y) is the region area of segmentation, coordinate (x, y) is the centroid of the area. norm(.) is a function that normalises the range of the value. Therefore, the 30-dim feature of the geometry composition is as follows

$$V_i = [\text{norm}(x_{i1}), \text{norm}(y_{i1}), \text{norm}(\text{area}(x_{i1}, y_{i1})), \ldots, \text{norm}(x_{i10}), \text{norm}(y_{i10}), \text{norm}(\text{area}(x_{i10}, y_{i10}))] \quad (3)$$

We set the non-existent regions by default parameters (x, y, area) = (0.5, 0.5, 0). The matching samples between the query and database images with geometry composition similarity are illustrated in Fig. 4.

4.3 Similarity matching-based templates

We need to find the most similar cluster as the composition templates based on the similarity matching. We search for the nearest neighbours of the query images in the 1024-dim low level visual feature space. Because of the high dimension, we choose to use the hierarchical K-means clustering to speed up the search process. We first divide the space into $K1$ clusters by K-means method ($K1 = 8$ in the experiment). Then, the query image is assigned to the closest cluster. We get the nearest results in the assigned cluster according to $D_{\text{att}}(I_i) = \text{dist}(U_q, U_i)$, which is the attention distance between the query image and photo in the dataset.

Fig. 5 visualises the matching process of attention composition feature. We use the Euclidean distance as the dist(.) to compute the similarity of the query images between the candidates and the cluster centre because of its effectiveness. The templates come from the clusters that are computed with the nearest distance for each image sequence. We collected the geometry and attention features from the cluster as the templates, reflecting the overall similarity of the corresponding professional candidates in terms of aesthetic composition.

5 Collaborative composition model

5.1 $\ell 2$ regularisation-based collaborative model

Inspired by the recent success of sparse representation and attribute construction in visual content understanding [22–26], we propose a collaborative composition model for representing the composition of query images according to the professional images. The idea from the tracking domain indicates that $F = [f_1, f_2, \ldots, f_n] \in \mathbb{R}^{d \times n}$ and $I = [i_1, i_2, ..., i_d] \in R^{d \times d}$ be the target templates and trivial templates, respectively. The linear representation can be written in matrix form as

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{\alpha}_F + \boldsymbol{I}\boldsymbol{\alpha}_I = [\boldsymbol{F}, \boldsymbol{I}]\begin{bmatrix}\boldsymbol{\alpha}_F \\ \boldsymbol{\alpha}_I\end{bmatrix} \doteq \boldsymbol{X}\boldsymbol{\alpha} \quad (4)$$

The candidate $\boldsymbol{y}$ is the tracking target, which should be lying in the subspace spanned by all target templates. Therefore, the representation coefficient vector is sparse and can be solved by $\ell 1$-norm minimisation

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \quad (5)$$

where λ is the regularisation parameter that controls the importance of the sparsity constrain to the reconstruction error. The weight of each target candidate can be computed as

$$w = \exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{F}\hat{\boldsymbol{\alpha}}\|_2^2}{\delta}\right) \quad (6)$$

where δ is a predefined parameter. In general, the candidate according to the biggest weight in w is considered as the target. To make the target image well approximated by a sparse linear combination of target templates with a small residual, we therefore assign a larger weight to the target image. To some extent, although $\ell 1$-norm regularisation can improve the representing performance by eliminating the effect of the unreasonable noise in the way of trivial templates, it still causes the activation of the trivial templates for non-sparse representation coefficients when representing the target image without occlusion. In addition, $\ell 1$-norm minimisation is very

IET Comput. Vis., 2015, Vol. 9, Iss. 5, pp. 691–698
© The Institution of Engineering and Technology 2015

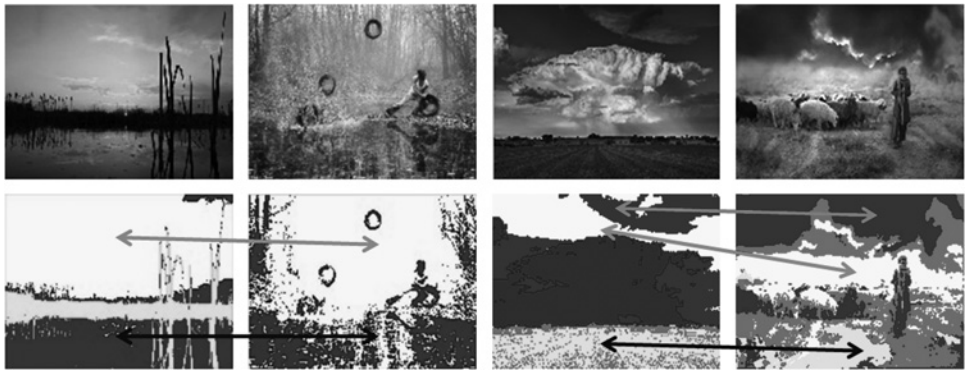

Fig. 4 *Matching samples of geometry composition features*

Parts with the same colour are matched

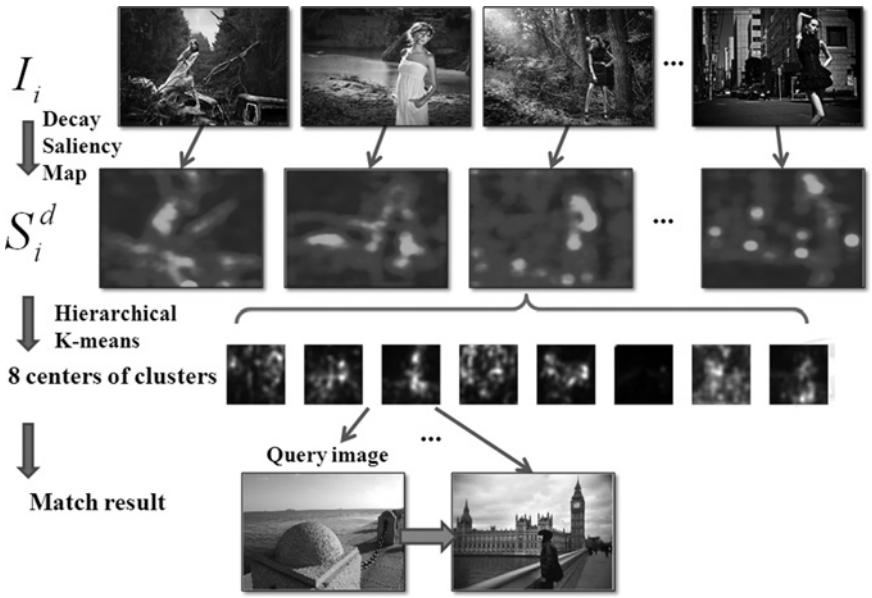

Fig. 5 *We collected the templates from the first stage cluster based on hierarchical K-means on attention features*

This stage is mainly to filter out most of the irrelevant photos according to the attention cue. Within the cluster, the next stage templates consist of both attention and geometry composition features

time-consuming with the total computational cost that grows proportionally as candidates increase. In [27], Xiao *et al.* proposed an efficient appearance representation-based tracker by adopting the $\ell 2$-regularised least square. Since the framework of $\ell 2$-norm regularisation is lack of sparsity constraint, most trivial templates will be activated, which the reconstruction errors of the target templates cannot be obtained accurately. We have confirmed that trivial templates are not suitable for the $\ell 2$-norm regularisation in the linear representation of the query images. In contrast, the combination of the attention and geometry composition templates contributes to discriminating the training templates set when a reasonable discriminative function based on their respective reconstruction error is designed.

Our work is inspired by this framework, in which the training photos are utilised as templates to improve the performance. This idea is based on the assumption that the query image can be better collaboratively represented by the training templates of aesthetic composition features. Thus a query image with a small reconstruction error when simultaneously represented by a set of certain attention and geometry templates indicates it is more likely to be recommended. The basic representation based on $\ell 2$-regularisation to calculate the coefficients is as follows

$$\boldsymbol{\gamma} = \arg\min_{\boldsymbol{\gamma}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_2 \tag{7}$$

The optimal solution of the above equation can be obtained via regularised least square as

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{8}$$

We can notice that the solution is obtained by taking the projection matrix $\boldsymbol{P} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}$ on candidate $\boldsymbol{y}$, in which the projection matrix $\boldsymbol{P}$ can be pre-calculated and independently from $\boldsymbol{y}$. By simply using $\boldsymbol{P}\boldsymbol{y}$ to obtain $\hat{\boldsymbol{\gamma}}$, it significantly reduces the computation cost.

The proposed collaborative composition model takes attention and geometry composition cues of the training images into account, thereby making the aesthetic representation more effective and robust. Initially, we extract both attention and geometry features for all the professional images (32×32 and 30 dim for the attention and geometry composition) for efficiency. In the first stage of clusters, we retrieve all the J attention templates in the nearest neighbourhood cluster by the hierarchical K-means as shown in Fig. 2. Within all the J templates, we continue processing them into $K2$ clusters. Thus, all the templates are stacked together to form the sets of attention templates in each second stage cluster C_i, $i \in [1, K2]$. In terms of reconstruction by the sets of templates, the training image set in each second-stage clusters C_i is composed of N_i attention templates (1024 dim)

© The Institution of Engineering and Technology 2015

and N_i geometry templates (30 dim) simultaneously. This means that the composition template set consists of both the templates of attention and geometry as $K2$ candidate dictionaries to get the reconstruction errors. This better facilitates photo recommendation containing the clusters that treated as the candidate reference dictionaries. As such, we use all the normalised features in each cluster C_i as the templates dictionary $\boldsymbol{D}_i$ and the feature query image is converted to a vector $\boldsymbol{y} \in \mathbb{R}^{G\times 1}$, where G denotes the size of the patch. The coefficient vector $\boldsymbol{\beta}$ of each query feature is computed by

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{D}_i\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2 \tag{9}$$

where the dictionary $\boldsymbol{D}_i \in \mathbb{R}^{G\times N_i}$ is generated from one of the $K2$ clusters containing N_i aesthetic composition feature in the cluster. λ is the weight parameter. In this work, the coefficient vector corresponding to each dictionary $\boldsymbol{D}_i$ is $\boldsymbol{\beta}_i \in \mathbb{R}^{N_i\times 1}$. All the $\boldsymbol{\beta}_i$ in the cluster is normalised and concatenated as

$$\mu = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{K2}] \in \mathbb{R}^{N_i\times K2} \tag{10}$$

μ is the responses related to all candidate dictionaries for one query feature. Next, we are trying to find the template set with the highest confidence score in μ using attention and geometry composition features collaboratively.

5.2 Confidence likelihood measurement

Specifically, we now assign the new notations for each feature templates. All attention template set $\boldsymbol{\mu}_A = [\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{K2}] \in \mathbb{R}^{N_i\times K2}$ is obtained in each second-stage cluster C_i, $i \in [1, K2]$ of the training templates, where N_i is the number of the attention templates. Subsequently, the weight of the attention template is updated based on the corresponding coefficient. When it is used to represent the candidate dictionaries, we select the one with the minimal reconstruction error. Similarly, we denote $\boldsymbol{\mu}_G = [\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_{K2}] \in \mathbb{R}^{N_i\times K2}$ as all geometry templates. We assume that the geometry templates only have the different sizes of feature with attention templates, that is, the other factors, such as the cluster numbers, scale of the templates are all the same with attention. In the following, we choose n candidates that correspond to the largest reconstruction errors as the reference templates. Let $\boldsymbol{y}$ be the query image, $\boldsymbol{y}_\mathrm{a}$ the attention feature vector and $\boldsymbol{y}_\mathrm{g}$ the geometry feature vector of the query image, which can be linearly represented by the attention and geometry templates set $\boldsymbol{A}_i = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{N_i}] \in \mathbb{R}^{1024\times N_i}$ $\boldsymbol{G}_i = [\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_{N_i}] \in \mathbb{R}^{30\times N_i}$, respectively. $\boldsymbol{y}_\mathrm{a}$ is linearly written as

$$\boldsymbol{y}_\mathrm{a} = u_1\boldsymbol{\alpha}_1 + u_1\boldsymbol{\alpha}_2 + \cdots + u_{N_i}\boldsymbol{\alpha}_{N_i} \tag{11}$$

At the same time, $\boldsymbol{y}_\mathrm{g}$ is linearly represented as

$$\boldsymbol{y}_\mathrm{g} = v_1\boldsymbol{g}_1 + v_1\boldsymbol{g}_2 + \ldots + v_{N_i}\boldsymbol{g}_{N_i} \tag{12}$$

As such, the optimal coefficient for each template set i can be calculated as

$$\hat{\boldsymbol{u}}_i = (\boldsymbol{A}_i^\mathrm{T}\boldsymbol{A}_i + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}_i^\mathrm{T}\boldsymbol{y}_\mathrm{a} \tag{13}$$

$$\hat{\boldsymbol{v}}_i = (\boldsymbol{G}_i^\mathrm{T}\boldsymbol{G}_i + \lambda\boldsymbol{I})^{-1}\boldsymbol{G}_i^\mathrm{T}\boldsymbol{y}_{\boldsymbol{g}} \tag{14}$$

After the templates are introduced into the dictionary, the weight of candidate can be computed as

$$w_i = \exp\left(-\frac{1}{1+\exp\left(-(\phi\varepsilon_{\mathrm{att}}^i + \varphi\varepsilon_{\mathrm{geo}}^i)/\sigma\right)}\right), \quad i \in K2 \tag{15}$$

where $\varepsilon_{\mathrm{att}}^i = \|\boldsymbol{y}_\mathrm{a} - \boldsymbol{A}_i\hat{\boldsymbol{u}}_i\|_2^2$ and $\varepsilon_{\mathrm{geo}}^i = \|\boldsymbol{y}_\mathrm{g} - \boldsymbol{G}_i\hat{\boldsymbol{v}}_i\|_2^2$ are the reconstruction errors of representing the candidate using attention and geometry templates, respectively. Constant σ is to balance the reconstruction errors $\varepsilon_{\mathrm{att}}$ and $\varepsilon_{\mathrm{geo}}$. In the implementation, we select the template dictionaries in each C_i only using the reconstruction error. Then, the candidate dictionaries according to the biggest weight are chosen from the clusters. The proposed method is developed based on the assumption that the composition of a query image can be better described by the joint representation in the subspace spanned by attention and geometry templates. Given a candidate image $\boldsymbol{y}$, it is represented by the training composition template set with the coefficients computed by the minimum value of w_i. Thus, we get the confidence value w_i of the candidate $\boldsymbol{y}$ within the template set C_i. Thus, it introduces discrimination in differentiating whether such images are recommended from both the attention and geometry perspectives. In addition, our confidence measure jointly exploits the distinct properties of the attention and geometry in computing the reconstruction errors to better distinguish suitable candidate set in the clusters.

5.3 Rules-based reference photo update

Adaptive reference photo selections based on aesthetic composition representation are presented as (15), where w_i denotes the candidates. In our experiment, we set $\phi = 0.6$, $\varphi = 0.4$, since different weights depend on personal preference. All the candidate photos which belong to the clusters with minimum w_i are selected. We associate the searched aesthetic candidates to the well-defined composition rules of professional photos. The rule of thirds and the golden ratio segmentation are the most well-known photograph composition principles. The ideas are to place the main subjects on the certain place of the photograph. Practically, we adapt four stress points ($S^1 \ldots S^4$) from rules of thirds and four stress points ($S^5 \ldots S^8$) from the golden ratio as well [3]. So we get eight points to determine the aesthetic appeal level of the visual subject centre. We measure the aesthetic score by the Euclidean distance between pose centre (x^0) and the eight stress points. The reference photo is to recommend with the minimum aesthetic score as

$$\begin{aligned} I_{\mathrm{ref}} &= \arg\min\left(\mathrm{Score}(I_{\mathrm{can}})\right) \\ \mathrm{Score} &= \min\left(\mathrm{dist}(x^0, s^i)\right), \quad i = 1, \ldots, 8 \end{aligned} \tag{16}$$

where I_{ref} is the selected reference photo. We evaluate the usefulness of collaborative composition model in the following section.

6 Experimental evaluation

We give extensive quantitative and qualitative results to evaluate the performance of our proposed method. The quantitative results compare our approach with several methods for satisfaction level in the user study on our collected dataset. To validate the effectiveness of the recommended results, we also investigate the qualitative performance of our recommended results.

6.1 Comparison methods and measurement

To the best of our knowledge, this is the first attempt on user pose recommendation in photography in terms of aesthetic consideration. Thus, we compare the proposed collaborative model with the baseline methods, which use different strategies to recommend the suitable poses and positions:

(1) Similarity-based computation method (similarity): To evaluate our collaborative composition model, we first compare our proposed method with the similarity-based computation method [28]. This method aims to recommend the suitable pose and position by taking the nearest neighbour distance to compute similarity, but ignoring the intrinsic relationship between the aesthetic features.
(2) Photo-quality assessment-based method (assessment): We also implemented the state-of-the-art photo-quality assessment approach

IET Comput. Vis., 2015, Vol. 9, Iss. 5, pp. 691–698
© The Institution of Engineering and Technology 2015

[15], which is modified for human pose recommendation. As described by Sun *et al.* [15], the human region usually has the high saliency in high quality photos. Similar with the process of similarity-based computation method, we first calculate the correspondences between the query and training images only in attention cue that is detected by the method [15] among the selected images, and obtain the best matched image as the recommended human position and pose.

6.2 Data collection

Most of the previous works on this topic collect private dataset to evaluate their performances. In order to guarantee the underlying professional spatial arrangements of the mined photos, we crawled photos from professional photo contest websites, namely myphoto (http://www.photo.net/) and DPchallenge (http://www.dpchallenge.com/). Our dataset consists of 232 photographs, mainly from editorial and fashion photographers' works. The test images are 50 images of scenes taken by cameras, mainly captured in scenes like campus, park, landscape and landmarks. We apply the 50 test images to the approaches for our experiments.

6.3 User study

The evaluations of the recommended photos based on different strategies are rather subjective without ground truths. We further design the following user study to measure the effectiveness of our proposed method. We conducted an effective evaluation of the proposed method as follows: the quality or aesthetics of a photo are related to the subjective evaluation. After getting the recommendation photos, ten participants (three female, seven male) from 23 to 30 years old help to give the evaluation of the poses and positions. Each measures the recommendation results of 50 images for five satisfactory levels. For the procedure of evaluation, each person is asked to mark the place and pose they want to capture in the images. Then, each user gives the satisfactory level by comparing with recommendation to make it subjective. There is no information about different strategies provided.

We propose a systematic criteria to measure the satisfaction levels, including quite satisfactory, satisfactory, relatively satisfactory, not too bad and dissatisfactory, which have the scores 5 to 1, respectively, for each photo. The participants are required to give a satisfaction score (i.e. 5 to 1) for a result. The average satisfaction score for each query image is collected to give quantitative results. For each method, we calculated the average score (AS) of human subject voting for each satisfaction level of feedback on all query images

$$\mathrm{AS}(r) = \frac{1}{N_p \times N_q} \sum l \times n_r^l \tag{17}$$

where $l = 1, 2, \ldots, 5$ denotes the users' satisfaction level, and n_r^l denotes the nth human subject voting for satisfactory level l for recommendation result r. N_p and N_q represent the total number of users and the number of query images.

The statistics of the user studies are illustrated in Fig. 6, in which the bars show the number of the proposed methods and the baselines in each satisfactory level, respectively. It is quite obvious that the recommendation results of our proposed method are more satisfactory than the baseline methods, while most results of the proposed collaborative composition model are above satisfactory.

Average satisfaction scores of different methods are shown in Table 1. The results show that the proposed model outperforms other methods with 5–23% improvement in the average satisfaction score. From Table 1, satisfaction level scores demonstrate that the proposed model is effective for photography recommendation, and is more satisfactory than other methods in users' subjective evaluation. Two reasons are behind the better performance of the proposed model. Firstly, the $\ell 2$-regularisation-based composition

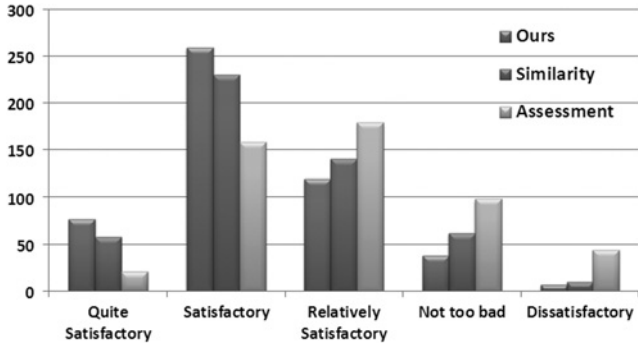

Fig. 6 *Statistics of the satisfaction level*

10 participants vote for 50 recommended results. We summarise the results of five levels ('quite satisfactory', 'satisfactory', 'relatively satisfactory', 'not too bad' and 'dissatisfactory') compared with baseline methods

Table 1 Average satisfaction score for each query images

Method	Ours	Similarity	Assessment
average score	3.72	3.52	3.03

model can collaboratively match attention and geometry in the reference images to improve the accuracy. In contrast, the similarity method ignores the view changes of query images, while our proposed model integrates the view information for effective estimating and predicting. Secondly, the learning strategy summarises the knowledge from the photo collection, while other methods only consider the composition information and are sensitive to the noise and condition changes. Compared with the assessment-based method, the improvement of our proposed method also demonstrates that our strategy is more reasonable and effective than single-cue low-level features based methods in solving the challenging semantic tasks. Overall, this study demonstrates that the collaborative composition model effectively solves this problem. However, some results of each method are still not that satisfactory, and it also demonstrates that the problem of photography recommendation is still a challenging and a hard problem.

6.4 Qualitative results analysis

We list three groups of examples (a), (b) and (c) to illustrate the qualitative results in Fig. 7, including several sample results of our method. In general cases, our method can give quite satisfying results, varying in lights, viewpoints and scale in (a) and (b). When the photo's attention composition is quite similar to the geometry composition in simple backgrounds, the recommendation results are quite satisfactory. Even through queries and the recommendation results are to some extent hardly related and with great visual changes, we still get some satisfying results. In such circumstances, the reference photo jointly depends on the attention and geometry composition similarities, which leads to promising performance.

We also list some failure results in (c), in which our method gives a dissatisfactory result, such as located in the corner of the image or on top of the river. This is because our method only considers more aesthetic factors of photos rather than semantics in the resulting images. Our matching method based on attention and geometry composition similarities still performs well in terms of capturing the composition features between the query and reference in these cases. The results of the poses and positions are difficult to achieve unless we take the spatial and semantic contextual information into consideration. These examples demonstrate that if we add semantic context into our model, we may obtain better results. While it requires high-level content analysis based on segmentation, we may explore more in the future work.

© The Institution of Engineering and Technology 2015

Fig. 7 *Experimental results of the recommendation pose and position by the proposed method*

In each case, left: query image; right: recommended photo. The annotated colour poses and positions are recommended from reference photos to input image

7 Conclusion

In this paper, we construct an aesthetic composition representation by utilising attention and geometry composition features. A novel pose and position recommendation strategy is proposed for portrait photography. The representation captures the scene structure of the photos and efficiently covers overall composition characteristics. By adopting the $\ell 2$-regularised least square, our collaborative composition model jointly considers the composition representation using the reconstruction error, which provides a fast performance to recommend the suitable reference candidates. We conduct adaptive reference photo selection with composition rules. It is worth noting that this model can be easily extended to more complex features and learning algorithms. User feedback is carried out on collected ground truth data. Experiments show that our model effectively recommends promising results.

8 Acknowledgments

This work was supported in part by the National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351802, and in part by the National Natural Science Foundation of China: 61133003, 61332016 and 61390510.

9 References

1 Cheng, B., Ni, B., Yan, S., Tian, Q.: 'Learning to photograph'. Proc. of ACM MM, 2010, pp. 291–300
2 Li, C., Gallagher, A., Loui, A.C., Chen, T.: 'Aesthetic quality assessment of consumer photos with faces'. IEEE ICIP, 2010, pp. 3221–3224
3 Obrador, P., Schmidt-Hackenberg, L., Oliver, N.: 'The role of image composition in image aesthetics'. IEEE ICIP, 2010, pp. 3185–3188
4 Eom, W., Lee, S., De Neve, W., Ro, Y.M.: 'Improving image tag recommendation using favorite image context'. IEEE ICIP, 2011, pp. 2445–2448
5 Li, C., Loui, A.C., Chen, T.: 'Towards aesthetics: a photo quality assessment and photo selection system'. Proc. of ACM MM, 2010, pp. 827–830
6 Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: 'Image quality assessment based on a degradation model', *IEEE Trans. Image Process.*, 2000, **9**, (4), pp. 636–650
7 Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: 'Image quality assessment: from error visibility to structural similarity', *IEEE Trans. Image Process.*, 2004, **13**, (4), pp. 600–612
8 Sheikh, H.R., Bovik, A.C., Cormack, L.: 'No-reference quality assessment using natural scene statistics: Jpeg2000', *IEEE Trans. Image Process.*, 2005, **14**, (11), pp. 1918–1927
9 Tong, H., Li, M., Zhang, H.-J., Zhang, C., He, J., Ma, W.-Y.: 'Learning no-reference quality metric by examples'. IEEE MMM, 2005, pp. 247–254
10 Freeman, M.: 'The complete guide to light and lighting in digital photography' (Sterling Publishing Company, Inc., 2007)
11 Ke, Y., Tang, X., Jing, F.: 'The design of high-level features for photo quality assessment'. IEEE CVPR, 2006, vol. 1, pp. 419–426
12 Tong, H., Li, M., Zhang, H.-J., He, J., Zhang, C.: 'Classification of digital photos taken by photographers or home users'. PCM, Springer, 2005, pp. 198–205
13 Datta, R., Joshi, D., Li, J., Wang, J.Z.: 'Studying aesthetics in photographic images using a computational approach'. ECCV, Springer, 2006, pp. 288–301
14 Luo, Y., Tang, X.: 'Photo and video quality evaluation: focusing on the subject'. ECCV, Springer, 2008, pp. 386–399
15 Sun, X., Yao, H., Ji, R., Liu, S.: 'Photo assessment based on computational visual attention model'. Proc. of ACM MM, 2009, pp. 541–544
16 Bhattacharya, S., Sukthankar, R., Shah, M.: 'A framework for photoquality assessment and enhancement based on visual aesthetics'. Proc. of ACM MM, 2010, pp. 271–280
17 Xu, P., Yao, H., Ji, R., Liu, X.-M., Sun, X.: 'Where should i stand? Learning based human position recommendation for mobile photographing', *Multimedia Tools Appl.*, 2014, **69**, (1), pp. 3–29
18 Yu, F.X., Ji, R., Chang, S.-F.: 'Active query sensing for mobile location search'. Proc. of ACM MM, 2011, pp. 3–12
19 Ferrari, V., Marin-Jimenez, M., Zisserman, A.: 'Pose search: retrieving people using their pose'. IEEE CVPR, 2009, pp. 1–8
20 Guo, C., Ma, Q., Zhang, L.: 'Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform'. IEEE CVPR, 2008, pp. 1–8
21 Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: 'Salient region detection and segmentation'. Computer Vision Systems, Springer, 2008, pp. 66–75
22 Zhang, S., Yao, H., Sun, X., *et al.*: 'Action recognition based on overcomplete independent components analysis', *Inf. Sci.*, 2014
23 Jiang, F., Zhang, S., Wu, S., Gao, Y., Zhao, D.: 'Multi-layered gesture recognition with kinect', *J. Mach. Learn. Res.*, 2014
24 Zhang, S., Yao, H., Sun, X., Lu, X.: 'Sparse coding based visual tracking: review and experimental comparison', *Pattern Recogn.*, 2013, **46**, (7), pp. 1772–1788
25 Zhang, Y., Qin, L., Yao, H., Xu, P., Huang, Q.: 'Beyond particle flow: bag of trajectory graphs for dense crowd event recognition'. IEEE ICIP, 2013
26 Zhang, Y., Qin, L., Ji, R., Yao, H., Huang, Q.: 'Social attribute aware force model: exploiting richness of interaction for abnormal crowd detection', *IEEE Trans. Circuits Syst. Video Technol.*, 2014
27 Xiao, Z., Lu, H., Wang, D.: 'Object tracking with l2-rls'. IEEE ICPR, 2012, pp. 1351–1354
28 Zhang, Y., Sun, X., Yao, H., Qin, L., Huang, Q.: 'Aesthetic composition representation for portrait photographing recommendation'. IEEE ICIP, 2012, pp. 2753–2756

© The Institution of Engineering and Technology 2015