1 Introduction

Next to lighting, posing is the most challenging aspect of portrait photography. A commonly adopted solution is to learn by example, which is beneficial for both trained photographers and novice users, especially when subjects have no clue about how to pose themselves. A collection of portrait images by professionals (e.g., [Perkins 2009]) provides a resource for photographers seeking inspiration for their own work. Such handful posing references (e.g., Posing App) have also been made available to smartphone platforms, which offer the unique possibility of directly overlaying camera view with a reference pose as visual guidance.

Using a collection of reference poses, however, is effective only if a desired reference pose can be quickly identified. For example, to beautify a given pose, a reference pose that is the most similar to the given pose is needed. A traditional way to search for a target pose is to let users manually browse the collection, which might be organized by category for faster access. It is obvious that the scalability of such completely manual browsing solutions is poor whereas the collection often keeps being expanded continuously given the limitless variations of pose. In fact there are literally thousands of ways to pose a subject [Smith 2004].

In this work we intend to provide more intuitive ways for photographers to get desired poses from a possibly large collection of reference poses. We present data-driven suggestions (Figure 1), which can serve as guidance and stimulate creativity, corresponding to two main operating modes in our system: refinement mode and exploration mode, respectively. The suggestions are automatically generated by real-time pose retrieval, using the current pose (i.e., for refinement mode) or a set of poses in existing portraits (i.e., for exploration mode) as a query. This is enabled by adopting consumption-level depth cameras (i.e., the Microsoft Kinect in our system as shown in Figure 1 (left)), which allow real-time and reasonably accurate estimation of the current pose. We will also show our preliminary results on context-based retrieval of reference poses relevant to the estimated scene semantics (see the accompanying video). Our work takes the first step of using consumer-level depth sensors towards more intelligent cameras for computational photography. Our preliminary experiments indicate that our tool is easy to use and useful for unskilled photographers.

2 Related Work

Data-driven suggestions for creativity support have been an active research topic very recently. Focusing on open-ended and creative discovery, various creativity support tools have been introduced for example in the context of freehand drawing [Lee et al. 2011] or 3D modeling [Chaudhuri and Koltun 2010]. Such ideas have also been applied to assisting non-professionals for taking more professional photos by making optimal composition suggestions [Cheng et al. 2010; Yin et al. 2012]. Our data-driven posing guide adds a new tool into the toolbox for creativity support.

The problem of pose retrieval, i.e., finding the most similar human poses to a given query pose, has been extensively studied in the fields of computer vision and graphics. The previous works can be mainly categorized into two groups: retrieving 2D poses of humans in images or videos (e.g., [Jammalamadaka et al. 2012]), and retrieving 3D poses from human motion data (e.g., [Choi et al. 2012]). The former often needs to address the problem of pose estimation while the latter typically employs temporal pose information, which is lacking in our problem. Our work is based on the success of pose retrieval in previous works, and focuses on a completely new application and its specific user interface design.
3 User Interface

We first introduce our system from the user’s perspective. Our discussions below focus on photographing individual subjects. Poses for men and women usually have their own characteristics. Therefore, for every new subject the photographer needs to specify the gender of the subject, unless the photographer is interested in applying female (male) poses to male (female) subjects. To capture a new picture with our tool, the photographer chooses one of two mutually exclusive suggestion modes: refinement mode and exploration mode, which will be explained in detail below. See Figure 2 for a screen shot of our main user interface in the refinement mode.

Refinement Mode. This mode allows the photographer to fine-tune the current pose of the subject with the help of one or more suggested reference poses (Figure 1 (a-c)). It is extremely beneficial when either the photographer or the subject has only a rough idea of a desired pose. To start with, the subject poses himself or herself, or is directed by the photographer to a rough pose in mind (Figure 1 (a)), possibly without any reference pose as a visual guide. Next the photographer may manually click the “search” button to get a set of suggested poses with respect to the current pose, or wait for our system to automatically return the relevant suggestions (Figure 1 (b)). The suggested poses are similar to the current pose but are generally aesthetically more pleasing.

Exploration Mode. It can be difficult to remain creative, especially for unskilled photographers. This mode serves as a creativity stimulation tool and is designed for photographers who are bored at the existing poses but are running out of pose ideas. To inspire the photographer, our tool provides him or her with a set of new poses (Figure 1 (e)) that are very dissimilar to the already taken poses (Figure 1 (d)). The current pose of the subject is less important in this mode (cf. the refinement mode). Before the photographer triggers the search process, he or she has to specify a directory where the existing portraits taken by our tool are stored. Otherwise, the poses already taken in the current session, each of which has its associated skeleton from the Kinect, is used as a query to find interesting poses. The newly taken pictures will be automatically added as part of a query for the next suggestion search (see the accompanying video).

No matter which mode is adopted, the photographer will be given a small set of suggestions, which may be either ignored or used by the photographer. To use the suggested poses, the photographer picks one reference pose appealing to him or her the most. Next the photographer uses the reference pose to guide the subject towards to a desired pose, which is usually a variant of the reference pose. Finally the photographer hits the software shutter button to take one shot. Please refer to the accompanying video for a live demo.

Context Option. Our system also supports a context-based search option, which can be used together with either the refinement or exploration model. Once the context option is enabled, our system returns only the suggestions that match the current context. For example, if a wall context is detected, only poses interacting with a wall will be suggested.

4 Methodology

Now we give the technical details on individual components in our system.

Dataset. Our system requires a large collection of professional portraits, from which relevant suggestions will be retrieved and provided to the photographer. The portraits in the dataset should be as diverse as possible and cover commonly seen poses (e.g., waist-up poses, three-quarter length reclining poses, three-quarter length seated poses, three-quarter length standing poses, full-length poses etc.). To enable the subsequent suggestion process, the underlying pose in each portrait should be recovered and represented by a standard skeleton which is compatible with that used by the Kinect (Figure 3).

To prepare such a dataset, we may download aesthetically pleasing portraits from the Internet and then recover the underlying pose in each portrait. Alternatively, we may hire professional photographers as well as professional models and capture various poses of the models using the Kinect with a roughly aligned high-resolution camera. While the second approach is able to provide much richer information such as 3D poses and 3D scene data, we take the first approach given millions of portraits publicly available on the Internet. Note that we are less interested in existing repositories of segmented human images like the one with PoseShop [Chen et al. 2012], since many pictures in such repositories are not professional portraits and not very suitable to serve as suggestions.

We found that the following 2.5D skeleton model works well in practice, though we are aware of semi-automatic approaches for recovering 3D human poses from monocular images e.g., [Agarwal and Triggs 2006; Taylor 2000]. With a simple drag-and-drop interface we first manually retarget a 15-joint skeleton model (Figure 3 (left)) used by the OpenNI framework to each collected portrait. Here, whether a certain joint appears in a dataset image or not is also manually specified. It is obvious that the depth information of a joint would be a discriminative feature for pose retrieval. To accommodate variations in body size, we take a relative depth metric \( d_{ij} \), which is defined as a signed angle in degree between the vector of joint \( i \) to joint \( j \) and the image plane, i.e., an out-of-plane angle \( d_{ij} \in [-90, 90] \). We provide a slider interface to specify \( d_{ij} \). In our implementation \( d_{ij} \) is defined for 11 bones, i.e., shoulder (DG; as illustrated in Figure 3), torso (BC), waist (MJ), and 8 limbs (DE, EF, GH, HI, JK, KL, MN, and NO). Lastly, for each portrait we manually assign relevant tags in different categories (e.g., gender: male/female, pose: standing/sitting/reclining, context: chair/wall/table/stairs).

Pose Matching. The formulation of distance metric for pose similarity which is compatible with human perception has been a fun-
We then define the distance between poses in a portrait image taken from a specific viewpoint. Let \( \theta \in [0, 360] \) be the oriented angle in degree from y-axis to a skeleton bone in the image plane, as illustrated in Figure 3. We represent a given pose (either a manually specified pose for a database image or a pose from the Kinect) as a 37-dimensional feature vector, denoted as \( P = (\theta_1, \ldots, \theta_{37}, d_1, \ldots, d_{12}) \), where \( \theta_i \) are the in-plane angles of the 14 regular bones in the skeleton and the 12 virtual bones (illustrated as dashed lines in Figure 3 (left)), and \( d_i \) are the out-of-plane angles representing the relative depth information.

The virtual bones are introduced for more robust matching. We then define the distance between poses \( P \) and \( Q \), i.e., pose dissimilarity, as follows:

\[
D(P, Q) = \sum_{i=1}^{26} \lambda_i^P f(\theta_i^P, \theta_i^Q) + \sum_{i=1}^{11} \lambda_i^d f(d_i^P, d_i^Q),
\]

where \( f(\alpha, \beta) = 1 - \cos(\alpha - \beta) \), \( \lambda_i^P \) and \( \lambda_i^d \) are the parameters to achieve semantically more meaningful matching results (a fixed set of parameter values used in our experiments). If a certain joint does not appear in either \( P \) or \( Q \), the corresponding terms in Equation 1 are ignored.

**Refinement Mode.** In this mode we seek dataset portraits whose corresponding poses are similar to the input pose from the Kinect, i.e., the current pose of the model, denoted as \( P_{\text{kinect}} \) (Figure 1 (a)). This is achieved by taking \( P_{\text{kinect}} \) as a query and retrieving the top-\( N \) (\( N = 20 \) in our experiments) matched poses from the dataset, i.e., looking for the dataset poses \( P \) with the minimum values of \( D(P_i, P_{\text{kinect}}) \). The returned list of suggestions might be very similar to each other. Our system allows the photographer to control the degree of diversity of suggestions by adopting the Maximal Marginal Relevance criterion, which is commonly used for a similar purpose in information retrieval [Chaudhuri and Koltun 2010].

**Exploration Mode.** This mode assumes the availability of a set of existing portraits (Figure 1 (d)), taken by our system and thus automatically coming with the corresponding poses from the Kinect. Let \( \{ Q_i \}_{1 \leq i \leq m} \) denote the set of poses in the available \( m \) portraits. To stimulate creativity, a suggested pose \( P \) should be very dissimilar to any of the poses in \( \{ Q_i \} \). On the other hand, it is preferable that the suggested pose can be achieved by modifying the current pose with ease. These two goals are formulated as follows:

\[
D_e(P) = \lambda^e D_r(P, P_{\text{kinect}}) - \lambda^d \min_{1 \leq i \leq m} D(P, Q_i),
\]

where \( D_r \) is a relaxed pose similarity metric defined over the thigh bones only, i.e., \( JK \) and \( MN \) as illustrated in Figure 3. Specifically, \( D_r(P, Q) = \lambda^d \sum_{i=1}^{12} f(d_i^P, d_i^Q) + \lambda^j \sum_{i=12}^{26} f(d_i^P, d_i^Q) \), where \( f(\alpha, \beta) \) is the same as that in Equation 1. We consider the thigh bones only, since significantly moving the thigh bones often leads to dramatic changes in pose, which should be largely avoided especially when capturing a sequence of shots [Pegram 2008]. To increase the diversity for every shot, the suggested poses are randomly picked from those that result in small values of \( D_e(P) \). In our experiments, we use the following fixed set of parameter values: \( \lambda^e = \lambda^j = 0.5 \) and \( \lambda^d = 0.8 \).

**Context Option.** Many of the poses in our repository are dependent on certain contexts or supporting objects. For example, the pose shown in Figure 4 (left) would be difficult to mimic without having a wall or similar supporting context. Thus it would greatly save the photographer’s time if our system is able to show only suggestions which are relevant to the current context. While this could be obtained by asking the photographer to manually pick one or more desired categories of poses, automatic identification of relevant suggestions to be presented to the user is definitely preferred.

We aim to provide active suggestions to the photographer. As a proof of concept, our current system only supports plane-like contexts such as the ground, walls, seating surfaces etc. We use a simple RANSAC-based approach to detect planar structures within the acquired depth map of the scene by the Kinect and then heuristically determine the types of the detected planes based on their properties such as orientation and area. However, it is well known that automatic high-level scene understanding is a challenging problem on its own and is not always robust even with more advanced algorithms like the one proposed by Nan et al [Nan et al. 2012]. Therefore our system only suggests context tags but lets the photographer confirm auto-tagging. See the accompanying video for live demos.

## 5 Pilot Study

Our current dataset contains 624 male and 1,185 female professional portraits, downloaded from Google Images and Flickr. We have done a pilot study, for which we recruited 10 participants, 3

![Figure 5: Representative results produced in the refinement mode. Our system allows photographers to easily refine the poses of subjects being photographed using the automatically suggestions as references. See the supplemental materials for more results.](image-url)
men and 7 women. None of them had received any professional training in photography and thus knew little about how to pose themselves gracefully. The participants came in pairs and each of them served as a photographer and a model in turn, i.e., alternating the roles of model and photographer. We thus had 10 unskilled photographers in total. To examine the refinement mode, the photographer was asked to guide the paired model to achieve 5 different poses in the photographer’s mind. To test the exploration mode, the photographer was required to capture another 5 poses in mind, which this time should be as diverse as possible and as dissimilar as possible to the previous 5 poses. Through a questionnaire-based survey at the end of the study, the participants confirmed that the suggested poses were useful (average score of 4.2 in a range where 1 is “strongly disagree” and 5 is “strongly agree”) and portrait posing with suggestions is easy to use (average score of 3.9). Figures 5 and 6 show the representative results in the refinement and exploration modes, respectively. Please refer to the supplemental materials for more results.

6 Future Work

First, we plan to conduct a user study to quantitatively evaluate the effectiveness of our tool by comparing the portraits taken with and without data-driven suggestions. Second, our current system focuses on photographing individual models. The proposed ideas could be easily extended to group portrait photography, where extra information such as the number of subjects and their relative height could be potentially incorporated for offering more constructive suggestions. Third, the recent technology from Extreme Reality enables real-time, software-only full body 3D motion capture, using a single standard 2D camera. It would be interesting to adopt such technologies and port our system to the popular smartphone platforms, making data-driven posing suggestions available to everyday users. Lastly, this work focuses on searching for desired poses or human positions given a specific point of view. Another closely related problem is the selection of the best view given a specific 3D arrangement of body parts and would be interesting to explore in the future.

Acknowledgements

We thank the reviewers for their constructive comments, the user study participants (namely, Victor Dibia, Audrey Samson, Guolei Zhang, Pengfei Xu, Qiaochu Mei, Qingkun Su, Xiaoyan Shen, Wanqi Li, Xiaozhu Zhang, Jiayue Yu, Xiaoyong Shen, Siyu Li) for their time, and Michael Brown for video narration. We are grateful to Flickr for allowing us to download so many images. This work was substantially supported by the Seed Grant from the City University of Hong Kong (No. 7003058).

References


Figure 6: The exploration mode allows photographers to easily explore new poses (b) that are very different from the existing poses (a). The portraits in (b) were created one by one (from left to right) with the newly created ones being added into the set of “existing” poses for deriving new suggestions. See the supplemental materials for more results.