Vis Comput
DOI 10.1007/s00371-013-0792-6

ORIGINAL ARTICLE

Internet visual media processing: a survey with graphics and vision applications

Shi-Min Hu · Tao Chen · Kun Xu · Ming-Ming Cheng · Ralph R. Martin

Abstract In recent years, the computer graphics and computer vision communities have devoted significant attention to research based on Internet visual media resources. The huge number of images and videos continually being uploaded by millions of people have stimulated a variety of visual media creation and editing applications, while also posing serious challenges of retrieval, organization, and utilization. This article surveys recent research as regards processing of large collections of images and video, including work on analysis, manipulation, and synthesis. It discusses the problems involved, and suggests possible future directions in this emerging research area.

Keywords Internet visual media · Large databases · Images · Video · Survey

1 Introduction

With the rapid development of applications and technologies built around the Internet, more and more images and videos are freely available on the Internet. We refer to these images and videos as *Internet visual media*, and they can be considered to form a large online database. This leads to opportunities to create various new data-driven applications, allowing non-professional users to easily create and edit visual media. However, most Internet visual media resources are unstructured and were not uploaded with applications in mind; furthermore, most are simply (and quite often inaccurately) indexed by text. Utilizing these resources poses a serious challenge. For example, if a user searches for ‘jumping dog’ using an Internet image search engine, the top-ranked results often contain results unrelated to those desired by the user who initiated the search. Some may contain a dog in a different pose to jumping, some may contain other animals which are jumping, some may contain cartoon dogs, and some might even contain a product whose brand name is ‘jumping dog’. Users have to carefully select amongst the many retrieved results, which is a tedious and time consuming task. Moreover, users expect most applications to provide interactive performance. While this may be straightforward to achieve for a small database of images or video, it becomes a problem for a large database.

A typical pipeline for Internet visual media processing consists of three steps: content retrieval, data organization and indexing, and data-driven applications. In the first step, meaningful objects are retrieved from chosen Internet visual media resources, for example by classifying the scene in each image or video, and extracting the contours of visually salient objects. This step can provide better labeling of visual media for content aware applications, and can compensate for the lack of accurate text labeling, as well as identifying the significant content. In the second step, the underlying relationships and correspondences between visual media resources are extracted at different scales, finding for example local feature similarities, providing object level classifications, determining object level similarities and dense correspondences, and so on up to scene level similarities. This information allows the construction of an efficient indexing and querying scheme for large visual media collections. For simplicity we will refer to this as database construction; it ensures that desired visual content can be rapidly retrieved. In the third step, Internet visual media applica-

S.-M. Hu (✉) · T. Chen · K. Xu · M.-M. Cheng
Tsinghua University, Beijing, China
e-mail: shimin@tsinghua.edu.cn

R.R. Martin
Cardiff University, Cardiff, UK

Published online: 22 March 2013

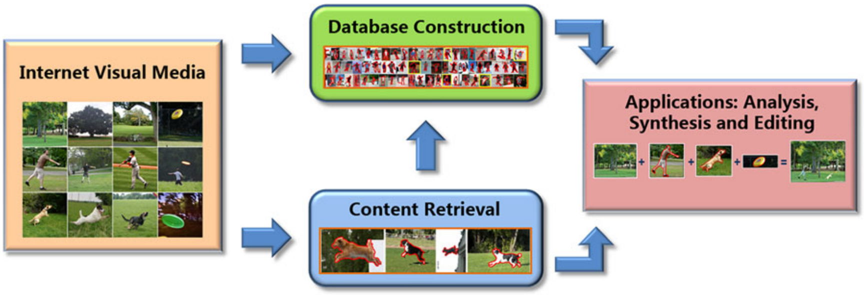

Fig. 1 A typical pipeline for Internet visual media processing

tions use these data. Traditional image and video processing approaches must be modified to adapt to this kind of data, and new approaches are also needed to underpin novel applications. The methods should be (i) *similarity aware*, to efficiently deal with the richness of Internet visual media: for example, a computation may be replaced by look-up of an image with similar appearance to the desired result, and (ii) *robust to variation*, to effectively deal with variations in visual media: semantically identical objects, e.g. dogs, can have widely varying appearances. Figure 1 indicates a typical pipeline for Internet visual media processing.

This survey is structured as follows: Sect. 2 introduces basic techniques and algorithms for Internet visual media retrieval. Section 3 describes recent work on visual media database construction, indexing, and organization. Sections 4–7 cover applications for visual media synthesis, editing, reconstruction, and recognition that utilize Internet visual media in unique ways. Section 8 draws conclusions, and discusses problems and possible future directions in this emerging area.

2 Internet visual media retrieval

This section describes approaches to Internet visual media retrieval, concentrating on content-based image retrieval methods. We start by introducing some basic techniques and algorithms that are key to Internet visual media retrieval.

2.1 Object extraction and matching

Extracting objects of interest from images is a fundamental step in many image processing tasks. Extracted objects or regions should be semantically meaningful primitives, and should be of high quality if they are to be used in applications like composition and editing. Object extraction can also reduce the amount of downstream processing in other applications, as it can then ignore other uninteresting parts of the image. A number of object extraction methods have been devised, both user assisted and automatic.

2.1.1 User-assisted object extraction

User interaction is an effective way of extracting high quality segmented objects. Such a strategy is adopted by most image editing software like Photoshop, and has the benefit of allowing users to modify the results when necessary. However, traditional approaches are tedious and time consuming: reducing the amount of user interaction while maintaining quality of the results is a key goal.

Early methods required users to precisely mark some parts or points on desired boundaries. An algorithm then searches for other boundary pieces which connect them with minimal cost, where costs depend on how unlikely it is that a location is part of the boundary. A typical example of such techniques is provided by *intelligent scissors* [82]. This method represents the image using a graph, and finds object boundaries by connecting user markers using Dijkstra's shortest path algorithm. However, this method has clear limitations—it requires precise marker placement, and has a tendency to take short cuts.

In order to relax the requirement of accurate marker placement, a number of methods have been developed to obtain an object boundary from a user-provided approximate initialization, such as *active contours* [63] and *level set methods* [93]. These methods produce an object segmentation boundary by optimizing an energy function which takes into account user preferences. However, energy functions used are often non-convex, so care is needed to avoid falling into incorrect local minimal. Doing so is nontrivial and methods often need re-initialization.

Providing robustness in the face of simple, inaccurate initialization, and permitting intuitive adjustment, are two trends in modern segmentation techniques. *Seeded image segmentation* [98, 125], e.g. *GraphCut* [10] and *random walks* [15, 42] are important concepts towards this aim. These methods allow users to provide a rough labeling of some pixels used as seeds while the algorithms automatically label other pixels; labels may indicate object or background, for example. Poor segmentation can be readily corrected by giving additional labels in wrongly labeled areas. Other techniques are also aimed at reducing the user annotation burden while still providing high quality results. *GrabCut* [89] successfully reduces the amount of user interaction to labeling a single rectangular region, by using an iterative version of GraphCut with a Gaussian mixture model (GMM) to capture color distribution. Bagon et al. [4] proposed a *segmentation-by-composition* approach which further reduces the required user interaction to selecting a single point.

Besides general interactive image segmentation, research has also considered particular issues such as high quality *soft segmentation* with transparent values in border regions [70, 105], or finding groups of similar objects [16, 121]. Typically, simpler user interaction may be provided at the cost of analyzing more complex attributes, e.g. *appearance distribution* [89], *symmetry* [4], and *self-similarity* [4, 16]. Such techniques are ultimately limited by the amount of computation that can be performed at interactive speed.

While a large variety of intuitive interactive tools exists for extracting objects of interest, and these can work well in some cases, these methods can still fail in more challenging cases, for example where the object has a similar color to the background, where the object has a complex texture, or where the boundary shape is complex or unusual in some way. This is also true of the automatic algorithms we consider next.

2.1.2 *Automatic object extraction*

Interactive methods are labor-intensive, especially in the context of the vast amount of Internet visual media. Therefore, automatic object extraction is necessary in many applications, such as *object recognition* [91], adaptive compression [19], *photo montage* [67], and *sketch based image retrieval* [13]. Extracting objects of interest, or *salient* objects, is an active topic in cognitive psychology [109], neurobiology [25], and computer vision [83]. Doing so enables various applications including visual media retargeting [45, 58, 112] as well as non-photorealistic rendering [47].

The human brain overcomes the problem of the huge amount of visual input by rapidly selecting a small subset of the stimuli for extensive processing [110]. Inspired by such human *attention mechanisms* [25, 65, 83, 109], a number of computational methods have been proposed to automatically find regions containing objects of interest in natural images. Itti et al. [57] were the first to devise a *saliency detection* method, which efficiently finds visually salient areas based on a center-surround model using multi-scale image features. Many further methods have been proposed to improve the performance of automatic salient object region detection and extraction [46, 78, 81]. Recently, Cheng et al. [17] proposed a global *contrast* based method for automatically extracting regions of objects in images, which has significantly better precision and recall than previous methods when evaluated on the largest publicly available dataset [1]. Automatically extracting salient objects enables huge numbers of Internet images to be processed, enabling many interesting image editing applications [13, 18, 55, 77].

2.1.3 *Object matching*

Object matching is an important problem in computer vision with applications to such tasks as object detection, image retrieval, and tracking. Most methods are based on matching *shape*. An early method by Hirata and Kato [51] expects a *precise* shape match, which is impractical in many real world applications. Alberto and Pala [24] employed an *elastic* matching method which improved robustness. Later, Belongie et al. [7] proposed a widely used *shape context* descriptor, which integrates global shape statistics with local point descriptors. Shape context is a discriminative tool for matching clean shapes on simple backgrounds, showing a 77 % retrieval rate on the challenging MPEG-7 dataset [92]. State-of-the-art shape matching approaches [6, 120] can achieve a 93 % retrieval rate on this dataset, by exploring similarities among a group of shapes. Although clean shapes can now be matched with an accuracy suitable for application use, shape matching in real world images is much more difficult due to cluttered backgrounds, and is still very challenging [107]. The major difficulty is to reliably segment objects of interest automatically. Various research efforts [5, 16] have been devoted to removing the influence of the background by selecting *clean* shape candidates for further matching, or modeling large scale locally translation invariant features [29].

2.2 Retrieval approaches

Content-based image retrieval (CBIR) [71] and *content-based video retrieval* [35] are popular research topics in statistics, pattern recognition, signal processing, multimedia, and computer vision. Many proposed CBIR systems are related to machine learning and often use *support vector machines* (SVM) for classification, based on color, texture, and shape as features. A comprehensive survey [23] covers this work. Much of it utilizes Internet visual media as the data source; in retrieval tasks; their methods are influenced by other kinds of retrieval such as Internet text retrieval. In 2003, Sivic et al. [99] presented *Video Google*, an object and scene retrieval approach that quickly retrieves a ranked list of key frames or shots from Internet videos for a given query video. It largely adapted text retrieval techniques, initially building a visual vocabulary of visual descriptors of all frames in the video database, as well as using inverted file systems and document rankings. Jing et al. [59] applied the PageRank idea [85] to large-scale image search, and the resulting *VisualRank* method proves that using content-based features can significantly improve image retrieval performance and generalize it to large-scale datasets.

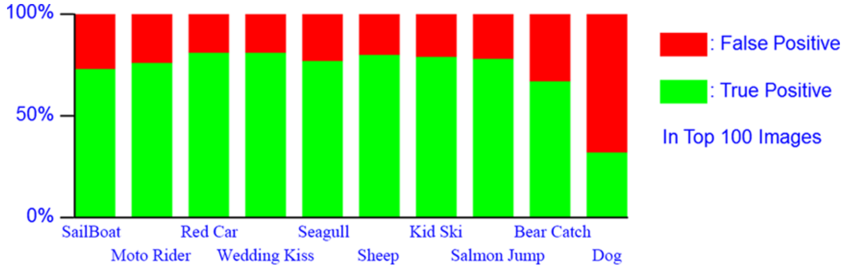

Fig. 2 Filtering performance for different scene item categories in *Sketch2Photo*

Other work proves that the richness of Internet visual media makes it possible to directly retrieve desired objects by filtering Internet images according to carefully chosen descriptors and strict criteria, providing intuitive retrieval tools for end users. *Sketch2Photo* [13] provides a very simple but effective approach to retrieval of well-segmented objects within Internet images based on user provided sketches and text labels. The key idea is to use an *algorithm-friendly* scheme for Internet image retrieval, which first discards images which have low usability in automatic computer vision analysis. Saliency detection and over-segmentation are used to reject images with complicated backgrounds. Shape and color consistency are used to further refine the set of candidate images. This retrieval approach can achieve a true positive rate of over 70 % amongst the top 100 ranked image objects. Figure 2 shows the filtering performance for various scene item categories.

Mindfinder [11] also has goals of efficient retrieval and intuitive user interaction when searching a database containing millions of images. Users sketch the main curves of the object to be found; they can also add text and coloring to further clarify their goals. Results are provided in real time by use of PCA hashing and inverted indices based on tags, color, and edge descriptors. However, the indexing method cannot deal with large differences in scale and position: in order to retrieve scenes the user intends, it must rely on the richness of the database.

Shrivastava et al. [95] present an image matching approach, which can find images which are visually similar at a higher level, even if quit different at the pixel level. Their method provides good *cross-domain* performance on the difficult task of matching images from different visual domains, such as photos and sketches, photos, and paintings, or photos under different lighting conditions. They use machine learning based on so-called *exemplar-SVM* to find the ‘data-driven uniqueness’ for each query image. Exemplar-SVM takes the query image as a positive training example and uses large numbers of Internet images as negative examples. Through a hard-negative mining process, it iteratively reveals the most unique feature in the form of a weighted Histogram of oriented gradients (HOG) descriptor [22] of the specific query image which is capable of cross-domain image matching.

Retrieval approaches utilizing large image databases or Internet images can be resistant to certain causes of failure in previous methods. *Sketch2Photo* [13], for example, can afford to discard a large number of true positives to overcome background complexity and ensure the quality of the information provided to later computer vision steps. In the work of [95], the negative examples may actually contain some true positives, but are swamped by the number of positive examples. The large scale and richness of the Internet data also allow these approaches to be more straightforward and practical. Severe filtering of data in these approaches, as in [13], can help to overcome the unreliable labeling of Internet visual media.

3 Visual media database construction and indexing

Internet visual media retrieval opens the door to various visual media applications. Many must respond to the user at interactive rates. However, the retrieval process is usually time consuming, especially for large-scale data. For example, the retrieval process in [13] takes 15 minutes to process a single image object (without taking into account image downloading time); training *exemplar-SVM* [95] for each query takes overnight on a cluster. One solution is to precompute important features of Internet visual media and to build a large visual media database with high level labeling and efficient indexing before querying commences, allowing the desired visual media to be retrieved within an acceptable time.

Researchers and industry have built many image databases for recognition and detection purposes. Some representative ones are the Yale *face database* [39], the INRIA database for *human detection* [22], *HumanEva* [96] for pose estimation, and the Daimler *pedestrian database* [31]. The image quality of these databases is usually limited, so they are not suitable for high quality synthesis and editing applications. Databases with better image quality like *PASCAL* VOC2006 [32], *Caltech 256* [43] and *LabelMe* [90] have no more than a few thousand images in each category, which is far from sufficient for Internet visual media tasks. Furthermore, these databases rarely contain accurate object contours, and manual segmentation is infeasible due to the workload.

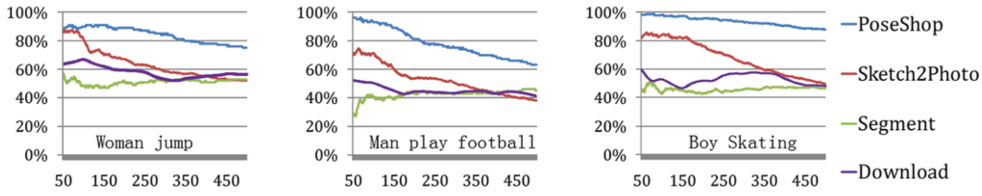

Fig. 3 Database quality vs. the highest ranked N images in *PoseShop*. *Horizontal axis*: number of (ranked) images in the constructed database. *Vertical axis*: true positive ratio of this method; it is consistently higher than for the method used in *Sketch2Photo*. This shows the advantage of using specific descriptors and detectors for particular object categories over using generic descriptors and saliency detectors

Although image resources on the Internet are vast, it is still very difficult to construct a content aware image database as very few of these images have content-based annotation. Various recent algorithms can automatically analyze images in large scale databases to discover specialised content such as *skies* [106] and *faces* [9]. Amazon's *Mechanical Turk* can also be helpful in such a process [104], but it should be carefully designed to optimize the division of labor between humans and machines. The recent *PoseShop* approach proposed by Chen et al. [14] allows almost fully automatic construction of a segmented human image database. Human action keywords were used to collect 3 million images from the Internet which were then filtered to be 'algorithm-friendly' [13], except that instead of saliency, specialized human, face, and skin detectors were used. The human contour is segmented, with cascade contour filtering used to the contours to match a few manually selected representative poses. Cloth attributes are also computed. The final database contains 400,000 well-segmented human pose images, organized by action, clothing, and pose in a tree structure. Queries can be based on silhouette sketches or skeletons. The accuracy of retrieval is demonstrated in Fig. 3.

Other approaches to efficiently query large visual media databases use database hashing. *ShadowDraw* [68] is a system for guiding freeform object drawing. The system generate a *shadow image* underlying the canvas in realtime when the user draws; the shadow image indicates to the user the best places to draw. The shadow image is composed from a large image database, by blending edge maps of retrieved candidate images automatically. To achieve realtime performance, it must efficiently match local edges in the current drawing with the database of images. This is achieved together with local and global consistency between the sketch and database images by use of a min-hash technique applied to a binary encoding of edge features. The *MindFinder* system also uses inverted file structures as indices to support real-time queries. To retrieve images of everyday products, Chung et al. [49] use a system based on a *bag of hash bits* (BoHB). This encodes local features of the image in a few hash bits, and represents the entire image by a bag of hash bits. The system is efficient enough for use on mobile devices.

Organizing large scale Internet image data at a higher level is also a difficult task. Heath et al. [50] construct *image webs* for such collections. These are graphs representing relationships between image pairs containing the same or similar objects. Image webs can be constructed by affine co-segmentation. To increase efficiency, two-phase co-segmentation candidate pairs are selected using CBIR and spectral graph theory. Analyzing image webs can produce visual hyperlinks and summary graphs for image collection exploration applications.

Overall, however, most organizing and indexing methods for Internet visual media databases concentrate on low level features and existing hashing techniques. A few attempts try to recover high level interconnections. Utilizing this information to build a database with greater semantic structure, and associated hashing algorithms are likely to become increasingly important topics.

4 Visual content synthesis

We now turn to various applications based on Internet visual media. The first kind uses them to create new content. Content synthesis and montage have long been desired by amateur users yet they are the privilege of the skilled professional, for whom they involve a great deal of tedious work. Various work has considered image synthesis for the purpose of generating artistic images [20, 53, 54], or realistic images, e.g. using alpha matting [56, 70, 105], gradient based blending [27, 87] or mean-value coordinates based blending [33, 123]. In recent years, data-driven approaches for image synthesis using large scale image databases or Internet images have provided promising improvements in ease of use and compositing quality.

Diakopoulos et al. [26] synthesized a new image from an existing image by replacing a user masked region marked with a *text label*. The masked area is filled using texture synthesis by finding a matching region from an image database containing manually segmented and labeled regions. However, extensive user interaction is still needed. Johnson et al. [60] generated a composite image from a group of *text*

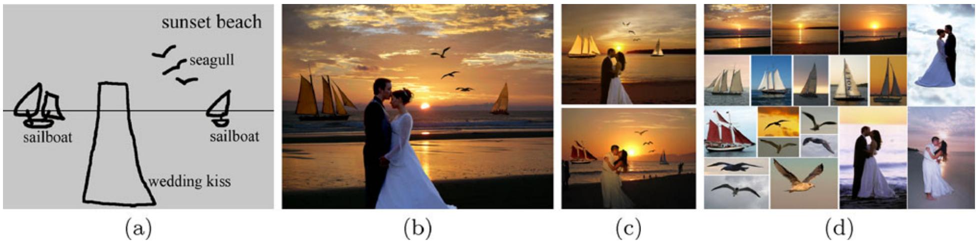

Fig. 4 Composition results of Sketch2Photo: (**a**) an input sketch plus corresponding text labels; (**b**) the composed picture; (**c**) two further compositions; (**d**) online images retrieved and used during composition. Image adapted from [13]

labels. The image is composited using graph cuts from multiple images retrieved from an automatic segmented and labeled image database, using the text labels. The method successfully reduces the time needed for user interaction, but has limited accuracy due to the limitations of automatic image labeling.

Lalonde et al. [67] presented the *photo clip art* system for inserting objects into an image. The user first specifies a class and a position for the inserted object, which is selected from a clip art database by matching various criteria including camera and lighting conditions. Huang et al. [55] converted an input image into an Arcimboldo-like stylized image. The input image is first segmented into regions using a mean-shift approach. Next, each region is replaced by a piece of *clip art* from a clip art database with the lowest matching cost. During matching, regions can be refined by merging or splitting to reduce the matching cost. However, this method is limited to a specific type of stylization. Johnson et al. [61] presented *CG2Real* to increase the realism of a rendered image by using real photographs. First, the rendered image is used to retrieve several similar real images from a large image database. Then, both real images and the rendered image are automatically co-segmented, and each region of the rendered image is enhanced using the real images by processes of color and texture transfer.

The *Sketch2Photo* system [13] combines *text labels* and *sketches* to select image candidates for Internet image montage. A hybrid image *blending* method is used, which combines the advantages of both Poisson blending and alpha blending, while minimizing the artifacts caused by these approaches. This blending operation has a unique feature designed for Internet images—it uses optimization to generate a numeric score for each blending operation at superpixel level, so it can efficiently compute an optimal combination from all the image candidates to provide the best composition results—see Fig. 4. Eitz et al. [30] proposed a similar *PhotoSketcher* system with the goal of synthesizing a new image only given user drawn *sketches*. The synthesized image is blended from several images, each found using a sketch. However, unlike *Sketch2Photo*, users must to put additionally scribbles on each retrieved image to segment the desired object.

Xue et al. [119] take a different data-driven approach to compositing fixed pairs of images. Statistics and visual perception were used together with an image database to determine which features have the most effect on realism in composite images. They used the results to devise a data-driven algorithm which can automatically alter the statistics of the source object so that it is more compatible with the target background image before compositing.

The above methods are limited to synthesis of a single image. To achieve multi-frame synthesis (e.g. for video), the main additional challenge is to maintain consistency of the same object across successive frames, since candidates for all frames usually cannot be found in the database. Chen et al. [14] proposed the *PoseShop* system, intended for synthesis of personalized images and multi-frame comic strips. It first builds a large human pose database by collecting and automatically segmenting online human images. Personalized Images or comic strips are produced by inputting a 3D human skeleton. Head and clothes swapping techniques are used to overcome the challenges of consistency, as well as to provide personalized results. See Fig. 5.

5 Visual media editing

The next application we consider is the editing of visual media. Manipulating and editing existing photographs is a very common requirement.

Various visual media editing tasks have been considered based on use of a few images or videos, such as tone editing [52, 76], colorization [69, 113, 114], edge-aware editing [12, 34, 117], dehazing [115, 122], detail enhancement [75, 86], edit propagation methods [3, 74, 118], and time-line editing [80]. Internet visual media can allow editing to become more intelligent and powerful by drawing on existing images and video. Much other similar work on data-driven image editing has been inspired by the well-known work of Hays and Efros [48] whose *image completion* algorithm is based on a very large online database of photographs. The algorithm completes holes in images using similar image regions from the database. No visible seams result, as similarity of scene images is determined by low level descriptors.

Traditional topics such as color enhancement, color replacement, and colorization can also benefit from the help

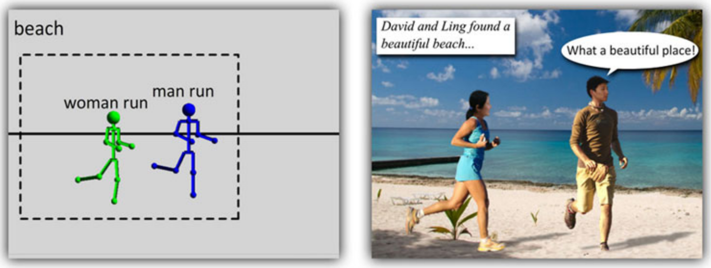

Fig. 5 PoseShop. *Left*: User provided skeletons (and some head images) for each character. *Right*: a generated photorealistic comic-strip. Image adapted from [14]

of Internet images. Liu et al. [79] proposed an example-based illumination independent colorization method. The input grayscale image is compared to several similar color images, obtained from the Internet as references. All are decomposed into an intrinsic component and an illumination component. The intrinsic component of the input image is colorized by analogy to the reference images, and combined with the illumination component of the input image to obtain the final result.

Wang et al. [111] proposed *color theme enhancement*, which is a data-driven method for recoloring an image which conveys a desired color theme. Color theme enhancement is formulated as an optimization problem which considers the chosen color theme, color constraints and prior knowledge obtained by analyzing an image database. Chia et al. [18] utilize Internet images for *colorization*. Using a similar framework to Sketch2Photo, the user can provide a text label and an approximate contour for the main foreground object in the image. The system then uses Internet images to find the best matches for color transfer. Multiple transfer results are generated and the desired result can be interactively selected using an intuitive interface.

Goldberg et al. [41] extend the framework of *Sketch2-Photo* to *image editing*. Objects in images can be interactively manipulated and edited with the help of similar objects in Internet images. The user provides text labels and an approximate contour of the object as in [18], and candidates are retrieved in a similar manner. The main challenge is to deform the retrieved object to agree with the selected object. Their novel alignment deformation algorithm optimizes the deformation both globally and locally along the object contour and inside the compositing seam. The framework supports various editing applications such as object completion, changes in occlusion relationships, and object augmentation and diminution. See Fig. 6.

6 3D scene reconstruction

A further class of applications centers around scene reconstruction. Many Internet images have captured the famous tourist attractions. Using these images to reconstruct 3D models and scenes of these attractions is a popular topic. Snavely [100] has summarized much of this work, so here we discuss some representative examples.

The *Photo Tourism* application of Snavely et al. [101] provides a 3D interface for interactively exploring a tourist scene through a multitude of photographs of the scene taken from differing viewpoints, with differing camera settings, etc. This method uses image-based modeling to reconstruct a sparse 3D model of the scene and to recover the viewpoint of each photo automatically for exploration. The exploration process can also provide smooth transitions between each viewpoint, with the help of image-based rendering. Further details are given in [103] of the structure-from-motion and image-based rendering algorithms used.

Again taking a large set of online photos as input, Snavely et al. [102] also considered how to find a *set of paths* linking interesting, i.e. frequently occurring, regions and viewpoints around such a site. These paths are then used to control image-based rendering, automatically computing sets of views all round a site, panoramas, and paths linking the most important views. These all lead to an improved browsing experience for the user.

Other researchers have considered even more ambitious goals, such as *reconstructing a whole city*. Agarwal et al. [2] achieved the traditionally impossible goal of building Rome in one day—their system uses parallel distributed matching and 3D reconstruction algorithms to process extremely large image collections. It can reconstruct a city from 150K images in under one day using a 500 core cluster. Frahm et al. [37] provided further advances in image clustering, stereo fusion and structure from motion to significantly improve performance. Their system can "build Rome in a cloudless day" using a single multicore PC with a modern graphics processor. Goesele et al. [40] and Furukawa et al. [38] both use multi-view stereo methods on large image data collections for *scene reconstruction*. The focus is on choice of appropriate subsets of images for matching in parallel.

Other problems tackled include the recovery of the *reflectance* of a static scene [44] to enable relighting purposes with the help of image collection, and an adaptation of photometric stereo for use on a large Internet-based set of images of a scene that computes both *global illumination* in each image and *surface orientation* at certain positions in the scene. This enables *scene summaries* to be generated [97],

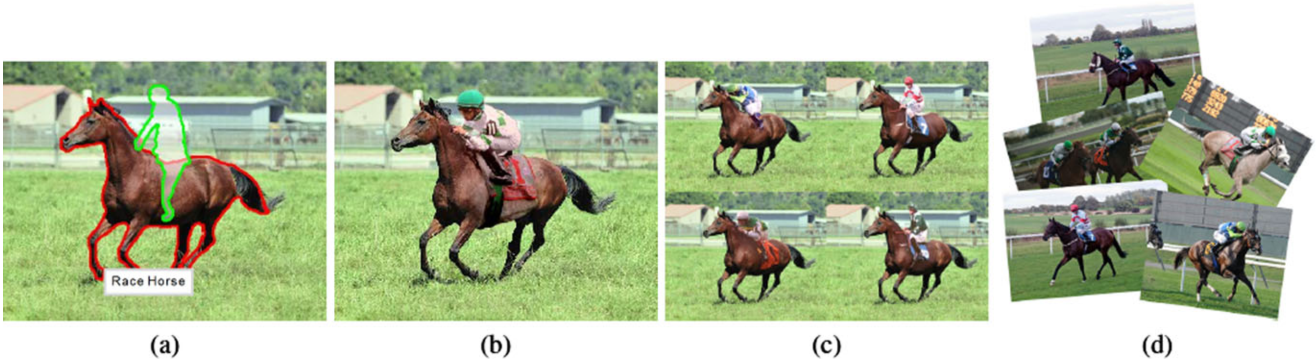

Fig. 6 Object-based image manipulation: (**a**) original image with object (horse) selected in red, and augmentation (jockey) sketched *in green*; (**b**) output image containing the added jockey and other regions of the source horse (saddle, reins, shadows); (**c**) alternative results (**d**) Internet photos used as sources. Image adapted from [41]

and more interestingly, the *weather* to be estimated in each image [94].

7 Visual media understanding

The final application area we consider is visual media understanding. In addition to specific applications like compositing, editing, and reconstruction, many Internet visual media applications aim to provide or make use of better understanding of visual media, either in a single item such as an image, or in a large set of related visual media, where the goal may be to identify a common property, or to summarize the media in some way, for example.

Various methods have advanced understanding of properties of a single image taken from a large image collection. Kennedy et al. [64] used Flickr to show that the metadata and tags of photograph can help improve the efficiency of vision algorithms. Kuthirummal et al. [66] showed how to determine *camera settings* and parameters for each image in a large set of online photographs captured in an uncontrolled manner, without knowledge of the cameras. Liu et al. [77] provide a way to determine the *best views* of 3D shapes based on images of them retrieved from the Internet.

Recently, much work has been done on image scene *recognition* and *classification* using visual, textual, temporal, and metadata information, especially geotagging of large image collections [21, 62, 72, 73, 88].

We first consider recognition. Quack et al. [88] consider recognizing objects from 200,000 community photos covering 700 square kilometers. They first cluster images according to visual, text, and spatial information. Then they use data-mining to assign appropriate text labels for each cluster, and further link it to a Wikipedia article. Finally verification is applied by checking consistency between the contents of the article and the assigned image cluster. The whole process is fully unsupervised and can be used for auto-annotation of new images. Li et al. [72] use 2D image features and 3D geometric constraints to summarize noisy Internet images of well-known landmarks such as the Statue of Liberty, and to build 3D models of them. The summarization and 3D models are then used for scene recognition for new images. Their method can distinguish and classify images in a database of different statues world-wide. Kalogerakis et al. [62] use geolocation for a sequence of photos with time-stamps. Using 6 million training images from Flickr, they train a prior distribution for travel that indicates the likelihood of travel endpoints in a certain time interval, and also the image likelihood for each location. The geolocation of a new image sequence is then decided by maximum likelihood methods. This approach can find accurate geolocation without recognizing the landmarks themselves. Zheng et al. [124] show how to *recognize* and *model sites* on a world-scale. Using an extensive list of such sites, they build visual models by image matching and clustering techniques. Their methods rely on English tour information websites; other languages are ignored, which limits recognition and model construction.

Turning to classification, Crandall et al. [21] use a mean shift approach to cluster the geotaggings harvested from 35 million photos in Flickr. This step can find the most popular places for people to take photos. Then this large photo collection is classified into these locations using a combination of textual, visual, and temporal features. Li et al. [73] studied landmark classification on a collection of 30 million images in which 2 million were labeled. They used multiclass SVM and bag-of-word SIFT features to achieve the classification. They also demonstrated that the temporal sequence of photos taken by a photographer used in conjunction with structured SVM allows greatly improved classification accuracy. These works shows the power of large scale databases for classification of photos of landmarks.

Other methods consider common properties of a set of images or attempt to summarize a set of images from a large image collection. In [72], the summarization and 3D modeling of a site of interest like the Statue of Liberty is based on an *iconic scene graph*, which consists of an iconic view of each image cluster clustered by GIST features [84], which is a low dimensional descriptor of a scene image. This iconic scene graph reveals the main aspects of the site and their geometric relationships. However, the process of finding iconic view discards many images and affects 3D modeling quality. Berg et al. [8] show how to automatically find *representative images* for chosen object categories. This could be very useful for large scale image data organization and exploration. Fiss et al. [36] consider how to *select still frames* from video for use as representative portraits. This work takes further steps towards modeling human perception, and could be extended to other subjects such as full body representative pose selection from sports video. Doersch et al. [28] show how to automatically retrieve most characteristic visual elements of a city, using a large geotagged image database. For example, certain balconies, windows, and street signs are characteristic of Paris. The method is currently limited to finding symbolic local elements of man made structures. Extending their method to large structures and natural elements would be an interesting attempt towards providing stylistic narratives of the whole world.

8 Conclusions

We have summarized recent research that organizes and utilizes large collections or databases of images and videos for purposes of assisting visual media analysis, manipulation, synthesis, reconstruction, and understanding. As already explained, most methods of organizing and indexing Internet visual media data are based on low level features and existing hashing techniques, with limited attempts to recover high level information and its interconnections. Utilizing such information to build databases with more semantic structure, and suitable indexing algorithms for such data are likely to become increasingly the focus of attention. The key to many applications lies in prepossessing visual media so that useful or interesting data can be rapidly found, and deciding what descriptors are best suited to a wide range of applications is an important open question.

Limitations of algorithmic efficiency also prevent full utilization of the huge amount of Internet visual media. Current methods work at most with millions of Internet images, which represent only a small portion. The more images that can be used, the better the results that can be expected. While parallelization will help, it is only part of the solution, and many core image processing techniques such as segmentation, feature extraction and classification are still potentially bottlenecks in any pipeline. Further work is needed on these topics.

There is a growing amount of work that tries to utilize all kinds of information in large scale data collections, not only visual information, but also metadata such as text tags, geotags, and temporal information. Event tags on images in social network web-sites are another potentially useful source of information, and in the longer term, it is possible that useful information can be inferred from the contextual information provided by such sites—for example, where a user lives or went on holiday can give clues to the content of a photo.

Finally, we note that work utilizing large collections of *video* is still scarce. Although it is natural to extend most image applications to video (see, for example, recent work that explores videos of famous scenes [108]), several reasons limit the ability to do this. Apart from the obvious limitation of processing time, many image processing and vision algorithms give unstable results when applied to video, or at least produce results with poor temporal coherence. Temporal coherence can be enforced in optimization frameworks, for example, but at the expense of even more computationally expensive procedures than processing the data frame by frame. Even state-of-the-art video object extraction methods may work well on some examples with minimum user interaction, but may fail if applied to a large collection of video data. Further, efficient, algorithms more specifically intended for Internet scale video collections are urgently needed, as exemplified by a recent paper on efficient video composition [116]. The idea of using 'algorithm-friendly' schemes to prune away videos that cannot be processed well automatically has yet to be applied.

Acknowledgements This work was supported by the National Basic Research Project of China (Project Number 2011CB302205), the Natural Science Foundation of China (Project Number 61120106007 and 61103079), and the National High Technology Research and Development Program of China (Project Number 2012AA011802), and an EPSRC Travel Grant.

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of IEEE CVPR, pp. 1597–1604 (2009)
2. Agarwal, S., Snavely, N., Simon, I., Seitz, S., Szeliski, R.: Building Rome in a day. In: Proceedings of IEEE ICCV, pp. 72–79 (2009)
3. An, X., Pellacini, F.: Appprop: all-pairs appearance-space edit propagation. ACM Trans. Graph. **27**(3), 40:1–40:9 (2008)
4. Bagon, S., Boiman, O., Irani, M.: What is a good image segment? A unified approach to segment extraction. In: Proceedings of ECCV, pp. 30–44. Springer, Berlin (2008)
5. Bai, X., Li, Q., Latecki, L., Liu, W., Tu, Z.: Shape band: a deformable object detection approach. In: Proceedings of IEEE CVPR, pp. 1335–1342 (2009)
6. Bai, X., Yang, X., Latecki, L.J., Liu, W., Tu, Z.: Learning context-sensitive shape similarity by graph transduction. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 861–874 (2010)
7. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002)

8. Berg, T., Berg, A.: Finding iconic images. In: Proceedings of IEEE CVPR, pp. 1–8 (2009)
9. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face swapping: automatically replacing faces in photographs. ACM Trans. Graph. **27**(3), 39:1–39:8 (2008)
10. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient n-d image segmentation. Int. J. Comput. Vis. **70**(2), 109–131 (2006)
11. Cao, Y., Wang, H., Wang, C., Li, Z., Zhang, L., Zhang, L.: Mindfinder: interactive sketch-based image search on millions of images. In: Proceedings of MM, pp. 1605–1608. ACM, New York (2010)
12. Chen, J., Paris, S., Durand, F.: Real-time edge-aware image processing with the bilateral grid. ACM Trans. Graph. **26**(3), 103:1–103:9 (2007)
13. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: Internet image montage. ACM Trans. Graph. **28**(5), 124:1–124:10 (2009)
14. Chen, T., Tan, P., Ma, L.Q., Cheng, M.M., Shamir, A., Hu, S.M.: Poseshop: human image database construction and personalized content synthesis. IEEE Trans. Vis. Comput. Graph. **19** (2013). http://doi.ieeecomputersociety.org/10.1109/TVCG.2012.148
15. Cheng, M.M., Zhang, G.X.: Connectedness of random walk segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 200–202 (2011)
16. Cheng, M.M., Zhang, F.L., Mitra, N.J., Huang, X., Hu, S.M.: Repfinder: finding approximately repeated scene elements for image editing. ACM Trans. Graph. **29**, 83:1–83:8 (2008)
17. Cheng, M.M., Zhang, G.X., Mitra, N., Huang, X., Hu, S.M.: Global contrast based salient region detection. In: Proceedings of IEEE CVPR, pp. 409–416 (2011)
18. Chia, A.Y.S., Zhuo, S., Gupta, R.K., Tai, Y.W., Cho, S.Y., Tan, P., Lin, S.: Semantic colorization with Internet images. ACM Trans. Graph. **30**(6), 156 (2011)
19. Christopoulos, C., Skodras, A., Ebrahimi, T.: The jpeg2000 still image coding system: an overview. IEEE Trans. Consum. Electron. **46**(4), 1103–1127 (2000)
20. Cong, L., Tong, R., Dong, J.: Selective image abstraction. Vis. Comput. **27**(3), 187–198 (2011)
21. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: Proceedings of WWW, pp. 761–770. ACM, New York (2009)
22. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of IEEE CVPR, vol. 1, pp. 886–893 (2005)
23. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: ideas, influences, and trends of the new age. ACM Comput. Surv. **40**(2), 5:1–50:60 (2008)
24. Del Bimbo, A., Pala, P.: Visual image retrieval by elastic matching of user sketches. IEEE Trans. Pattern Anal. Mach. Intell. **19**(2), 121–132 (1997)
25. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. **18**(1), 193–222 (1995)
26. Diakopoulos, N., Essa, I., Jain, R.: Content based image synthesis. In: Proceedings of CIVR, pp. 299–307 (2004)
27. Ding, M., Tong, R.F.: Content-aware copying and pasting in images. Vis. Comput. **26**(6–8), 721–729 (2010)
28. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes Paris look like Paris? ACM Trans. Graph. **31**(4), 101:1–101:9 (2012)
29. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: benchmark and bag-of-features descriptors. IEEE Trans. Vis. Comput. Graph. **17**(11), 1624–1636 (2011)
30. Eitz, M., Richter, R., Hildebrand, K., Boubekeur, T., Alexa, M.: Photosketcher: interactive sketch-based image synthesis. IEEE Comput. Graph. Appl. **31**(6), 56–66 (2011)
31. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: survey and experiments. IEEE Trans. Pattern Anal. Mach. Intell. **31**(12), 2179–2195 (2009)
32. Everingham, M., Zisserman, A., Williams, C.K.I., van Gool, L.: The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results (2006). http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf
33. Farbman, Z., Hoffer, G., Lipman, Y., Cohen-Or, D., Lischinski, D.: Coordinates for instant image cloning. ACM Trans. Graph. **28**(3), 67:1–67:9 (2009)
34. Fattal, R., Carroll, R., Agrawala, M.: Edge-based image coarsening. ACM Trans. Graph. **29**(1), 6 (2009)
35. Feng, B., Cao, J., Bao, X., Bao, L., Zhang, Y., Lin, S., Yun, X.: Graph-based multi-space semantic correlation propagation for video retrieval. Vis. Comput. **27**(1), 21–34 (2011)
36. Fiss, J., Agarwala, A., Curless, B.: Candid portrait selection from video. ACM Trans. Graph. **30**(6), 128:1–128:8 (2011)
37. Frahm, J.M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.H., Dunn, E., Clipp, B., Lazebnik, S., Pollefeys, M.: Building Rome on a cloudless day. In: Proceedings of ECCV, pp. 368–381. Springer, Berlin (2010)
38. Furukawa, Y., Curless, B., Seitz, S., Szeliski, R.: Towards Internet-scale multi-view stereo. In: Proceedings of IEEE CVPR, pp. 1434–1441 (2010)
39. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 643–660 (2001)
40. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.: Multi-view stereo for community photo collections. In: Proceedings of IEEE ICCV, pp. 1–8 (2007)
41. Goldberg, C., Chen, T., Zhang, F.L., Shamir, A., Hu, S.M.: Data-driven object manipulation in images. Comput. Graph. Forum **31**(2pt1), 265–274 (2012)
42. Grady, L.: Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **28**(11), 1768–1783 (2006)
43. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical report (2007)
44. Haber, T., Fuchs, C., Bekaer, P., Seidel, H.P., Goesele, M., Lensch, H.: Relighting objects from image collections. In: Proceedings of IEEE CVPR, pp. 627–634 (2009)
45. Han, D., Sonka, M., Bayouth, J.E., Wu, X.: Optimal multiple-seams search for image resizing with smoothness and shape prior. Vis. Comput. **26**(6–8), 749–759 (2010)
46. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Proceedings of NIPS, Cambridge, MA, pp. 545–552 (2006)
47. Hata, M., Toyoura, M., Mao, X.: Automatic generation of accentuated pencil drawing with saliency map and LIC. Vis. Comput. **28**(6–8), 657–668 (2012)
48. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Trans. Graph. **26**(3), 4:1–4:7 (2007)
49. He, J., Feng, J., Liu, X., Cheng, T., Lin, T.H., Chung, H., Chang, S.F.: Mobile product search with bag of hash bits and boundary reranking. In: Proceedings of IEEE CVPR, pp. 3005–3012 (2012)
50. Heath, K., Gelfand, N., Ovsjanikov, M., Aanjaneya, M., Guibas, L.: Image webs: computing and exploiting connectivity in image collections. In: Proceedings of IEEE CVPR, pp. 3432–3439 (2010)
51. Hirata, K., Kato, T.: Query by visual example - content based image retrieval. In: Proceedings of EDBT' 92, pp. 56–71. Springer, London (1992)
52. Huang, H., Xiao, X.: Example-based contrast enhancement by gradient mapping. Vis. Comput. **26**(6–8), 731–738 (2010)
53. Huang, H., Zang, Y., Li, C.F.: Example-based painting guided by color features. Vis. Comput. **26**(6–8), 933–942 (2010)

54. Huang, H., Fu, T.N., Li, C.F.: Painterly rendering with content-dependent natural paint strokes. Vis. Comput. **27**(9), 861–871 (2011)
55. Huang, H., Zhang, L., Zhang, H.C.: Arcimboldo-like collage using Internet images. ACM Trans. Graph. **30**(6), 155:1–155:8 (2011)
56. Huang, M.C., Liu, F., Wu, E.: A gpu-based matting Laplacian solver for high resolution image matting. Vis. Comput. **26**(6–8), 943–950 (2010)
57. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
58. Jin, Y., Liu, L., Wu, Q.: Nonhomogeneous scaling optimization for realtime image resizing. Vis. Comput. **26**(6–8), 769–778 (2010)
59. Jing, Y., Baluja, S.: Visualrank: applying pagerank to large-scale image search. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1877–1890 (2008)
60. Johnson, M., Brostow, G.J., Shotton, J., Arandjelović, O., Kwatra, V., Cipolla, R.: Semantic photo synthesis. Comput. Graph. Forum **25**(3), 407–413 (2006)
61. Johnson, M.K., Dale, K., Avidan, S., Pfister, H., Freeman, W.T., Matusik, W.: Cg2real: improving the realism of computer generated images using a large collection of photographs. IEEE Trans. Vis. Comput. Graph. **17**(9), 1273–1285 (2011)
62. Kalogerakis, E., Vesselova, O., Hays, J., Efros, A., Hertzmann, A.: Image sequence geolocation with human travel priors. In: Proceedings of IEEE ICCV, pp. 253–260 (2009)
63. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. Int. J. Comput. Vis. **1**(4), 321–331 (1988)
64. Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How flickr helps us make sense of the world: context and content in community-contributed media collections. In: Proceedings of MM, pp. 631–640. ACM, New York (2007)
65. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. **4**(4), 219–227 (1985)
66. Kuthirummal, S., Agarwala, A., Goldman, D.B., Nayar, S.K.: Priors for large photo collections and what they reveal about cameras. In: Proceedings of ECCV, pp. 74–87. Springer, Berlin (2008)
67. Lalonde, J.F., Hoiem, D., Efros, A.A., Rother, C., Winn, J., Criminisi, A.: Photo clip art. ACM Trans. Graph. **26**(3), 3:1–3:10 (2007)
68. Lee, Y.J., Zitnick, C.L., Cohen, M.F.: Shadowdraw: real-time user guidance for freehand drawing. ACM Trans. Graph. **30**, 27:1–27:10 (2011)
69. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. ACM Trans. Graph. **23**(3), 689–694 (2004)
70. Levin, A., Lischinski, D., Weiss, Y.: A closed-form solution to natural image matting. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 228–242 (2008)
71. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM Trans. Multimed. Comput. Commun. Appl. **2**(1), 1–19 (2006)
72. Li, X., Wu, C., Zach, C., Lazebnik, S., Frahm, J.M.: Modeling and recognition of landmark image collections using iconic scene graphs. In: Proceedings of ECCV, pp. 427–440. Springer, Berlin (2008)
73. Li, Y., Crandall, D., Huttenlocher, D.: Landmark classification in large-scale image collections. In: Proceedings of IEEE ICCV, pp. 1957–1964 (2009)
74. Li, Y., Ju, T., Hu, S.M.: Instant propagation of sparse edits on images and videos. Comput. Graph. Forum **29**(7), 2049–2054 (2010)
75. Ling, Y., Yan, C., Liu, C., Wang, X., Li, H.: Adaptive tone-preserved image detail enhancement. Vis. Comput. **28**(6–8), 733–742 (2012)
76. Lischinski, D., Farbman, Z., Uyttendaele, M., Szeliski, R.: Interactive local adjustment of tonal values. ACM Trans. Graph. **25**(3), 646–653 (2006)
77. Liu, H., Zhang, L., Huang, H.: Web-image driven best views of 3d shapes. Vis. Comput. **28**(3), 279–287 (2012)
78. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 353–367 (2011)
79. Liu, X., Wan, L., Qu, Y., Wong, T.T., Lin, S., Leung, C.S., Heng, P.A.: Intrinsic colorization. ACM Trans. Graph. **27**(5), 152:1–152:9 (2008)
80. Lu, S.P., Zhang, S.H., Wei, J., Hu, S.M., Martin, R.R.: Time-line editing of objects in video. IEEE Trans. Vis. Comput. Graph. **19** (2013). http://doi.ieeecomputersociety.org/10.1109/TVCG.2012.145
81. Ma, Y.F., Zhang, H.J.: Contrast-based image attention analysis by using fuzzy growing. In: Proceedings of MM, pp. 374–381. ACM, New York (2003)
82. Mortensen, E.N., Barrett, W.A.: Intelligent scissors for image composition. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques SIGGRAPH '95, pp. 191–198. ACM, New York (1995)
83. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: Proceedings of IEEE CVPR, vol. 2, pp. 2049–2056 (2006)
84. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
85. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web (1999)
86. Pajak, D., Cadík, M., Aydin, T.O., Okabe, M., Myszkowski, K., Seidel, H.P.: Contrast prescription for multiscale image editing. Vis. Comput. **26**(6–8), 739–748 (2010)
87. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Trans. Graph. **22**(3), 313–318 (2003)
88. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In: Proceedings of CIVR, pp. 47–56. ACM, New York (2008)
89. Rother, C., Kolmogorov, V., Blake, A.: "Grabcut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23**(3), 309–314 (2004)
90. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. Int. J. Comput. Vis. **77**(1–3), 157–173 (2008)
91. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition? In: Proceedings of IEEE CVPR, vol. 2, pp. 37–44 (2004)
92. Salembier, P., Sikora, T., Manjunath, B.: Introduction to MPEG-7: Multimedia Content Description Interface. Wiley, New York (2002)
93. Sethian, J.: Level set methods and fast marching methods: evolving interfaces. In: Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science, vol. 3. Cambridge University Press, Cambridge (1999)
94. Shen, L., Tan, P.: Photometric stereo and weather estimation using Internet images. In: Proceedings of IEEE CVPR, pp. 1850–1857 (2009)
95. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. ACM Trans. Graph. **30**(6), 154:1–154:10 (2011)
96. Sigal, L., Black, M.: Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Brown University Technical Report 120 (2006)

97. Simon, I., Snavely, N., Seitz, S.: Scene summarization for online image collections. In: Proceedings of IEEE ICCV, pp. 1–8 (2007)
98. Sinop, A., Grady, L.: A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm. In: Proceedings of IEEE ICCV, pp. 1–8 (2007)
99. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of IEEE ICCV, vol. 2, pp. 1470–1477 (2003)
100. Snavely, N.: Scene reconstruction and visualization from Internet photo collections: a survey. IPSJ Trans. Comput. Vis. Appl. **3**(0), 44–66 (2011)
101. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM Trans. Graph. **25**(3), 835–846 (2006)
102. Snavely, N., Garg, R., Seitz, S.M., Szeliski, R.: Finding paths through the world's photos. ACM Trans. Graph. **27**(3), 15:1–15:11 (2008)
103. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from Internet photo collections. Int. J. Comput. Vis. **80**(2), 189–210 (2008)
104. Sorokin, A., Forsyth, D.: Utility data annotation with amazon mechanical turk. In: Proceedings of IEEE CVPR, pp. 1–8 (2008)
105. Sun, J., Jia, J., Tang, C.K., Shum, H.Y.: Poisson matting. ACM Trans. Graph. **23**(3), 315–321 (2004)
106. Tao, L., Yuan, L., Sun, J.: Skyfinder: attribute-based sky image search. ACM Trans. Graph. **28**(3), 68:1–68:5 (2009)
107. Thayananthan, A., Stenger, B., Torr, P., Cipolla, R.: Shape context and chamfer matching in cluttered scenes. In: Proceedings of IEEE CVPR, vol. 1, pp. 127–133 (2003)
108. Tompkin, J., Kim, K.I., Kautz, J., Theobalt, C.: Videoscapes: exploring sparse, unstructured video collections. ACM Trans. Graph. **31**(5), 68:1–68:12 (2012)
109. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cogn. Psychol. **12**(1), 97–136 (1980)
110. Tsotsos, J.: Analyzing vision at the complexity level. Behav. Brain Sci. **13**(3), 423–469 (1990)
111. Wang, B., Yu, Y., Wong, T.T., Chen, C., Xu, Y.Q.: Data-driven image color theme enhancement. ACM Trans. Graph. **29**(6), 146:1–146:10 (2010)
112. Wang, D., Li, G., Jia, W., Luo, X.: Saliency-driven scaling optimization for image retargeting. Vis. Comput. **27**(9), 853–860 (2011)
113. Welsh, T., Ashikhmin, M., Mueller, K.: Transferring color to greyscale images. ACM Trans. Graph. **21**(3), 277–280 (2002)
114. Wu, J., Shen, X., Liu, L.: Interactive two-scale color-to-gray. Vis. Comput. **28**(6–8), 723–731 (2012)
115. Xiao, C., Gan, J.: Fast image dehazing using guided joint bilateral filter. Vis. Comput. **28**(6–8), 713–721 (2012)
116. Xie, Z.F., Shen, Y., Ma, L.Z., Chen, Z.H.: Seamless video composition using optimized mean-value cloning. Vis. Comput. **26**(6–8), 1123–1134 (2010)
117. Xie, Z.F., Lau, R.W., Gui, Y., Chen, M.G., Ma, L.Z.: A gradient-domain-based edge-preserving sharpen filter. Vis. Comput. **28**(12), 1195–1207 (2012)
118. Xu, K., Li, Y., Ju, T., Hu, S.M., Liu, T.Q.: Efficient affinity-based edit propagation using k-d tree. ACM Trans. Graph. **28**(5), 118:1–118:6 (2009)
119. Xue, S., Agarwala, A., Dorsey, J., Rushmeier, H.: Understanding and improving the realism of image composites. ACM Trans. Graph. **31**(4), 84:1–84:10 (2012)
120. Yang, X., Koknar-Tezel, S., Latecki, L.: Locally constrained diffusion process on locally densified distance spaces with applications to shape retrieval. In: Proceedings of IEEE CVPR, pp. 357–364 (2009)
121. Zhang, F.L., Cheng, M.M., Jia, J., Hu, S.M.: Imageadmixture: putting together dissimilar objects from groups. IEEE Trans. Vis. Comput. Graph. **18**(11), 1849–1857 (2012)
122. Zhang, J., Li, L., Zhang, Y., Yang, G., Cao, X., Sun, J.: Video dehazing with spatial and temporal coherence. Vis. Comput. **27**(6–8), 749–757 (2011)
123. Zhang, Y., Tong, R.: Environment-sensitive cloning in images. Vis. Comput. **27**(6–8), 739–748 (2011)
124. Zheng, Y.T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.S., Neven, H.: Tour the world: building a web-scale landmark recognition engine. In: Proceedings of IEEE CVPR, pp. 1085–1092 (2009)
125. Zhong, F., Qin, X., Peng, Q.: Robust image segmentation against complex color distribution. Vis. Comput. **27**(6–8), 707–716 (2011)

Shi-Min Hu received the Ph.D. degree from Zhejiang University in 1996. He is currently a professor in the Department of Computer Science and Technology at Tsinghua University, Beijing. His research interests include digital geometry processing, video processing, rendering, computer animation, and computer-aided geometric design. He is associate Editor-in-Chief of *The Visual Computer* (Springer), and on the editorial boards of *Computer-Aided Design* and *Computer & Graphics* (Elsevier). He is a member of the IEEE and ACM.

Tao Chen is a postdoctoral researcher at Department of Computer Science and Technology in Tsinghua University. He received his Ph.D. degree from the same university in 2011. His research interests include: image/video processing, editing and composition.

Kun Xu is an assistant professor at Department of Computer Science and Technology in Tsinghua University. Before that, He received his Ph.D. degree from the same university in 2009. His research interests include: real-time rendering, appearance modeling, image/video editing.

Ming-Ming Cheng is a research fellow at Oxford Brookes University. Earlier, he received his Ph.D. degree from Tsinghua University in 2012 under guidance of Prof. Shi-Min Hu. His research primarily centers on algorithmic issues in image understanding and processing.

Ralph R. Martin is currently a Professor at Cardiff University. He obtained his Ph.D. degree in 1983 from Cambridge University. He has published more than 250 papers and 13 books, covering such topics as solid and surface modeling, intelligent sketch input, geometric reasoning, reverse engineering, and various aspects of computer graphics. He is a Fellow of: the Learned Society of Wales, the Institute of Mathematics and its Applications, and the British Computer Society. He is on the editorial boards of *Computer Aided Design*, *Computer Aided Geometric Design*, *Geometric Models*, the *International Journal of Shape Modeling*, *CAD and Applications*, and the *International Journal of CADCAM*.