Flickr Distance: A Motion Prediction Approach for Visual Concepts

T.Suganya
P.G Scholar
SNS college of Engineering

B. Anuradha
Assistant Professor
SNS college of Engineering

B.Chellaprabha
Assistant Professor
SNS college of Engineering

Abstract
While image alignment has been studied in different areas of computer vision for decades, aligning images depicting different scenes remains a challenging problem. Analogous to optical flow where an image is aligned to its temporally adjacent frame, we propose SIFT flow, a method to align an image to its nearest neighbors in a large image corpus containing a variety of scenes. The SIFT flow algorithm consists of matching densely sampled; pixel-wise SIFT features between two images, while preserving spatial discontinuities. The SIFT features allow robust matching across different scene/object appearances, whereas the discontinuity- preserving spatial model allows matching of objects located at different parts of the scene. The proposed approach will robustly aligns complex scene pairs by determining Flickr Distance between the image concepts. The Flickr distance between two concepts is defined as the Jensen-Shannon (J-S) divergence between their LTVLM. Based on SIFT flow, we propose an alignment- based large database framework for image analysis and synthesis, where image information is transferred from the nearest neighbors to a query image according to the dense scene correspondence. This framework can be demonstrated through concrete applications, such as motion field prediction from a single image, motion synthesis via object transfer, satellite image registration and face recognition.

Index Terms
Artificial Intelligence, Image Analysis, Distance Learning, Machine Vision, Scene alignment, SIFT flow, motion prediction for a single image, motion synthesis via object transfer.

1. INTRODUCTION

Similarity measurement has been studied for decades and it remains a hot research topic especially in multimedia literature. There are quite a few important applications related to this measurement, including indexing, ranking, clustering, annotation, and recommendation. Conceptual correlation measurement calculates the semantic distance between these two concepts and determines its Flickr Distance [5]. Image alignment, registration and correspondence are central topics in computer vision. There are several levels of scenarios in which image alignment dwells. The simplest level, aligning different views of the same scene, has been studied for the purpose of image stitching and stereo matching. The image alignment problem becomes more complicated for dynamic scenes in video sequences, e.g. optical flow estimation [7]. The correspondence between two adjacent frames in a video is often formulated as an estimation of a 2D flow field. In this work, we are interested in a new, higher level of image alignment: aligning two images from different 3D scenes but sharing similar scene characteristics. Image alignment at the scene level is thus called scene alignment. The two images to match may contain object instances captured from different viewpoints, placed at different spatial locations, or imaged at different scales. The two images may also contain different quantities of objects of the same category, and some objects present in one image might be missing in the other. Due to these issues the scene alignment problem is extremely challenging. Inspired by optical flow methods, which are able to produce dense, pixel-to-pixel correspondences between two images, we propose SIFT flow [1], adopting the computational framework of optical flow, but by matching SIFT descriptors instead of raw pixels. In SIFT flow, a SIFT descriptor [8] is extracted at each pixel to characterize local image structures and encode contextual information. A discrete, discontinuity preserving, flow estimation algorithm is used to match the SIFT descriptors between two images. The use of SIFT features allows robust matching across different scene/object appearances and the spatial model allows matching of objects located at different parts of the scene. Moreover, a coarse-to-fine matching scheme is designed to significantly accelerate the flow estimation process.

2. RELATED WORK

Image alignment, image registration or correspondence, is a broad topic in computer vision, computer graphics and medical imaging, covering stereo, motion analysis, video compression, shape registration, and object recognition. It is beyond the scope of this paper to give a thorough review on image alignment.. In this section, we will review the image alignment literature focusing on

Manuscript received February 5, 2014
Manuscript revised February 20, 2014

(a) What to align, or the features that are consistent across images, e.g. pixels, edges, descriptors;
(b) Which way to align, or the representation of the alignment, e.g. sparse vs. dense, parametric vs. nonparametric;
(c) How to align, or the computational methods to obtain alignment parameters.
In addition, correspondence can be established between two images, or between an image and image models. In image alignment we must first define the features based on which image correspondence will be established: an image measurement that does not change from one image to another. In stereo and optical flow the brightness constancy assumption was often made for building the correspondence between two images. But researchers came to realize that pixel values are not reliable for image matching due to changes of lighting, perspective and noise . Features such as phase , filter banks , mutual information and gradient are used to match images since they are more reliable than pixel values across frames, but they still fail to deal with drastic changes. Middle-level representations such as scale-invariant feature transform (SIFT) [8], shape context [9], histogram of oriented gradients (HOG) [1] have been introduced to account for stronger appearance changes, and are proven to be effective in a variety of applications such as visual tracking , optical flow estimation and object recognition .
The representation of the correspondence is another important aspect of image alignment. One can utilize the information of every pixel to obtain a dense correspondence, or merely use sparse feature points. The form of the correspondence can be pixel-wise displacement such as a 1-D disparity map (stereo) and a 2-D flow field (optical flow), or parametric models such as affine and homography. Although a parametric model can be estimated from matching every pixel and a dense correspondence can be interpolated from sparse matching typically, pixel-wise displacement is obtained through pixel- wise correspondence, and parametric motion is estimated from sparse, interest point detection and matching. In between the sparse and dense representation is correspondence on contours, which has been used in tracking objects and analyzing motion for texture less objects. The fact that the underlying correspondence between scenes is complicated and unclear, and detecting contours from scenes can be unreliable, leads us to seek for dense, pixel-wise correspondence for scene alignment. Estimating dense correspondence between two images is a nontrivial problem with spatial regularity, i.e. the displacements (flow vectors) of neighboring pixels tend to be similar.
When the feature values of the two images are close and temporally smooth, this displacement can be formulated as a continuous variable and the estimation problem is often reduced to solving PDE's using Euler-Lagrange. When the feature values are different, or other information such as occlusion needs to be taken into account, one can use belief propagation and graph cuts to optimize objective functions formulated on Markov random fields. A dual-layer formulation is proposed to apply tree-reweighted BP to estimate optical flow fields. These advances in inference allow us to solve dense scene matching problems effectively. Image representations, such as color histograms, texture models, segmented regions GIST descriptors, bag of words and spatial pyramids have been proposed to find similar images at a global level. Common to all these representations is the lack of meaningful correspondences across different image regions, and therefore, spatial structural information of images [3] tends to be ignored. Our interest is to establish dense correspondences between images across scenes, an alignment problem that can be more challenging than aligning images from the same scene and aligning images of the same object category since we wish all the elements that compose the scene to be aligned. Our work relates to the task of co-segmentation [2] that tried to simultaneously segment the common parts of an image pair, and to the problem of shape matching [9] that was used in the context of object recognition.

Inspired by the recent advances in image alignment and scene parsing, we propose SIFT flow to establish the correspondence between images across scenes.. In this paper, we will explore the Flickr Distance and SIFT flow algorithm in more depth and will demonstrate a wide array of applications for SIFT flow.

3. OUR CONTRIBUTION

A collection of images for a concept are taken from a source using object identification mechanism. Each concept may consist of different views and sequences for images. Effective image alignment based on the object and its sequence pattern, is a complicated one hence the distance between the two objects is calculated first and based on that, prediction and identification of next objects is done. The proposed approach uses Flickr Distance (FD) for measuring the conceptual relations between the images and the Histogram methodology for robustly identifying and aligning complex scene pairs containing significant spatial difference. An alignment based large image database for image analysis and synthesis is constructed, where image information is transferred from the nearest neighbours to a query image according to the distance.The proposed work will be useful in the areas of motion field prediction, pattern analysis, pattern synthesis from a single static image, image registration and object recognition.

4. FLICKR DISTANCE

As mentioned before, FD is based on representing a concept by building a statistical model from a set of related images, and the concept distance is defined by the distance of the two corresponding models. The framework of calculating Flickr distance is shown in Fig. 1. Therefore there are two basic problems in this scheme.First, an expansive image data set that can reflect most of the concepts relationships well. Second, a reasonable statistical model that can capture the visual relationships of the concepts well.

4.1 Visual Concept Pool

To simulate the concurrence of concepts in human cognition, the calculation of conceptual correlation should be performed in daily life environment. To achieve this, we try to mine the statistical semantic relations between concepts from a large pool of the daily life photos. To obtain a less- biased estimation, the image pool should be very large and the source of the images should be independent. Luckily, the online photo sharing website Flickr meets both conditions. There are more than 109 photos on Flickr, and these photos are uploaded by independent users. Besides, each photo has a crowd of manual tags, which provide good connections between the photos and the semantic concepts (tags). Thus, it is an ideal data set for learning the visual conceptual relations.

A set of images related to the concept are collected by the tag-based retrieval on Flickr. LTVLM is adopted to model the images. The conceptual correlation which is used to connect the concepts is measured by some distance measurements, such as the Jensen-Shannon divergence

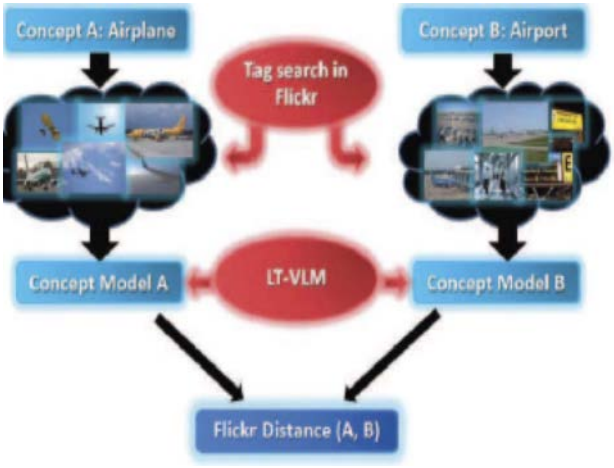

Fig. 1 Framework for measuring the conceptual distance.

4.2 Visual Model Selection

To analyze the conceptual correlation in a large Flickr photo pool, visual language model (VLM), an efficient visual statistical analysis method, is adopted.VLM is more discriminative than the well-known bag-of-words (BoW) model. Superior to BoW, VLM captures not only local appearance features but also their spatial dependence, which is more discriminative in characterizing the concept than the pure visual feature distribution. The training of VLM is fast, which makes the modeling method especially suitable for large scale conceptual data set. The output of VLM is conditional distributions of visual features, based on which a strict distance metric can be easily defined.

4.3 Concept Modeling

In this, we elaborate on the concept modeling process. A single visual word does not correspond to the specific semantic meaning due to the semantic gap problem. That is one of the open problems in computer vision. We assume the spatial correlation between the words contains some information. We use the language model, which models the trigrams of the visual words. One visual word may have multiple meanings, while the meaning of a trigram has a higher probability of being unique.

5. FEATURE EXTRACTION

5.1 The Sift Flow Algorithm

SIFT is a local descriptor to characterize local gradient information [8]. In [8], SIFT descriptor is a sparse feature representation that consists of both feature extraction and detection. In this paper, however, we only use the feature extraction component. For every pixel in an image, we divide its neighborhood (e.g. 16x16) into a 4x4 cell array, quantize the orientation into 8 bins in each cell, and obtain a 4x4x8 = 128- dimensional vector as the SIFT representation for a pixel. We call this per-pixel SIFT descriptor SIFT image.

To visualize SIFT images, we compute the top three principal components of SIFT descriptors from a set of images, and then map these principal components to the principal components of the RGB space via projection of a 128D SIFT descriptor to a 3D subspace, we are able to compute the SIFT image from an RGB image . In this visualization, the pixels that have similar color may imply that they share similar local image structures. Note that this projection is only for visualization; in SIFT flow, the entire 128 dimensions are used for matching.

5.2 Matching Objective

We designed an objective function similar to that of optical flow to estimate SIFT flow from two SIFT images. Similar to optical flow, we want SIFT descriptors to be matched along the flow vectors, and the flow field to be smooth, with discontinuities agreeing with object boundaries. Based on these two criteria, the objective function of SIFT flow is formulated as follows. Let p = (x, y) be the grid coordinate of images, and w(p) = (u(p), v(p)) be the flow vector at p. We only allow u(p) and v(p) to be integers and we assume that there are L possible states for u(p) and v(p), respectively. Let s1 and s2 be two SIFT images that we want to match. Set ε contains all the spatial neighborhoods (a four-neighbor system is used). The energy function for SIFT flow is defined as:

$$E(W) = \sum_p \min(\| s_1(p) - s_2(p + W(p)) \|_1, t) + \quad (1)$$

$$\sum_p \eta (|u(p)| + |v(p)|) + \quad (2)$$

$$\sum_{(p,q)\in\varepsilon} \min(\alpha|u(p) - u(q)|, d) + \min(\alpha|v(p) - v(q)|, d) \quad (3)$$

which contains a data term, small displacement term and smoothness term. The data term in Eqn.1 constrains the SIFT descriptors to be matched along with the flow vector w(p). The small displacement term inEqn. 2 constrains the flow vectors to be as small as possible when no other information is available. The smoothness term in Eqn. 3 constrains the flow vectors of adjacent pixels to be similar. In this objective function, truncated L1 norms are used in both the data term and the smoothness term to account for matching outliers and flow discontinuities, with t and d as the threshold, respectively.

5.3 Neighborhood of SIFT flow

In theory, we can apply optical flow to two arbitrary images to estimate a correspondence, but we may not get a meaningful correspondence if the two images are from different scene categories. In fact, even when we apply optical flow to two adjacent frames in a video sequence, we assume dense sampling in time so that there is significant overlap between two neighboring frames. Similarly, in SIFT flow, we define the neighborhood of an image as the nearest neighbors when we query a large database with the input. Ideally, if the database is large and dense enough to contain almost every possible image in the world, the nearest neighbors will be close to the query image, sharing similar local structures.

5.4 Coarse-to-fine matching scheme

Despite the speed up, directly optimizing Eqn. (3) using this dual-layer belief propagation scales poorly with respect to image dimension. In SIFT flow, a pixel in one image can literally match to any pixels in the other image. Suppose the image has h2 pixels, then L ≈ h, and the time and space complexity of this dual-layer BP is O(h4). For example, the computation time for 145x105 images with an 80x80 search window is 50 seconds. It would require more than two hours to process a pair of 256x256 images with a memory usage of 16GB to store the data term. To address the performance drawback, we designed a coarse-to-fine SIFT flow matching scheme that significantly improves the performance. The basic idea is to roughly in Eqn. estimate the flow at a coarse level of image grid, then gradually propagate and refine the flow from coarse to fine. The procedure is illustrated in Figure 6. For simplicity, we use s to represent both s1 and s2. A SIFT pyramid{s(k)} is established, where s(1) = s and s(k+1) is smoothed and downsampled from s(k). At each pyramid level k, let pk be the coordinate of the pixel to match, ck be the offset or centroid of the searching window, and w(pk) be the best match from BP. At the top pyramid level s(3), the searching window is centered at p3 (c3 = p3) with size mxm, where m is the width (height) of s(3). The complexity of BP at this level is O(m4). After BP converges, the system propagates the optimized flow vector w(p3) to the next (finer) level to be c2 where the searching window of p2 is centered. The size of this searching window is fixed to be nxn with n = 11. This procedure iterates from s(3) to s(1) until the flow vector w(p1) is estimated. The complexity of this coarse-to-fine algorithm is O(h2 log h), a significant speed up compared to O(h4). Moreover, we double η and retain α and d as the algorithm moves to a higher level of pyramid in the energy minimization.

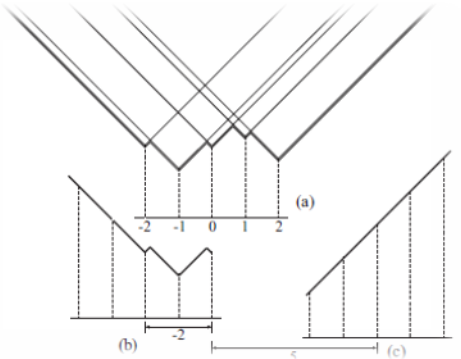

Fig. 7. We generalized the distance transform function for truncated L1 norm to pass messages between neighboring nodes that have different offsets (centroids) of the searching window.

When the matching is propagated from an coarser level to a finer level, the searching windows for two neighboring pixels may have different offsets (centroids). We modify the distance transform function developed for truncated L1 norm to cope with this situation, with the idea illustrated in Figure 7. To compute the message passing from pixel p to its neighbor q, we first gather all other messages and data term, and apply the routine in to compute the message from p to q assuming that q and p have the same offset and range. The function is then extended to be outside the range by increasing α per step, as shown in Figure 7 (a). We take the function in the range that q is relative to p as the message. For example, if the offset of the searching window for p is 0, and the offset for q is 5, then the message from p to q is plotted in Figure 7 (c). If the offset of the searching window for q is –2 otherwise, the message is shown in Figure 7 (b). Using the proposed coarse-to-fine matching scheme and modified distance transform function, the matching between two 256x256 images takes 31 seconds on a workstation with two quad-core 2.67 GHz Intel Xeon CPUs and 32 GB memory, in a C++ implementation. Further speedup (up to 50x) can be achieved through GPU implementation of the BP-S algorithm since this algorithm can be parallelized.

A natural question is whether the coarse-to-fine matching scheme can achieve the same minimum energy as the ordinary matching scheme (using only one level). We randomly selected 200 pairs of images to estimate SIFT flow, and check the minimum energy obtained using coarse-to-fine scheme and ordinary scheme (non coarse-to-fine), respectively. For these 256x256 images, the average running time of coarse-to- fine SIFT flow is 31 seconds, compared to 127 minutes in average for the ordinary matching. The coarse-to-fine scheme not only runs significantly faster, but also achieves lower energies most of the time compared to the ordinary matching algorithm as shown in Figure 8.

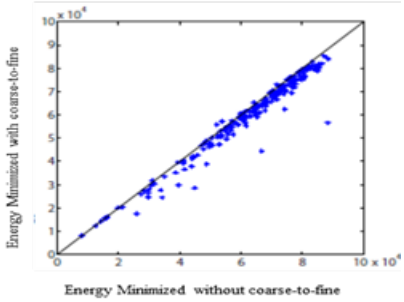

Fig. 8. Coarse-to-fine SIFT flow not only runs significantly faster, but also achieves lower energies most of the time.

6.CONCLUSION

In this paper, we propose the Flickr Distance to measure conceptual distance. The visual characteristic of the concepts is modeled by a novel latent topic visual language model. J-S divergence between the LTVLM can be deemed as a measurement of the conceptual distance. Both subjective user study and objective experiment show that Flickr distance is more coherent to human cognition than NGD and TCD and we introduced the concept of dense scene alignment: to estimate the dense correspondence between images across scenes. We proposed SIFT flow to match salient local image structures with spatial regularities, and conjectured that match- ing in a large database using SIFT flow leads to semantically meaningful correspondences for scene alignment. We further proposed an alignment-based large database framework for image analysis and synthesis, where image information is transferred from the nearest neighbors in a large database to a query image according to the dense scene correspondence estimated by SIFT flow. This framework is concretely realized in motion prediction from a single image, motion synthesis via object transfer and face recognition.

REFERENCES

[1] Ce Liu, Member IEEE, Jenny Yuen, Member IEEE, and Antonio Torralba, Member IEEE (2011) - "SIFT Flow: Dense Correspondence across Scenes and its Applications"

[2] Dizan Vasquez & Thierry Fraichard Inria Rhône-Alpes & Lab. Gravir, Grenoble (FR) - "Motion Prediction for Moving Objects: A Statistical Approach".

[3] Haipeng Zhang, Mohammed Korayem, Erkang You, David.J (2012) - "Beyond Co-occurrence :Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities".

[4] Josef Sivic and Andrew Zisserman Robotics Research Group, United Kingdom (2003) - "Video Google: A Text Retrieval Approach to Object Matching in Videos".

[5] Lei Wu, Member, IEEE, Xian-Sheng Hua, Member, IEEE, Nenghai Yu, Member, IEEE, Wei-Ying Ma and Shipeng Li (2012) - "Flickr Distance: A Relationship Measure for Visual Concepts".

[6] Zhiyu Zhou , Jianxin Zhang, Li Fang, Zhejiang Sci - China (2009) - " Object Tracking Based on Dynamic Template and Motion Prediction."

[7] B. K. P. Horn and B. G. Schunck. Determing optical flow. Artificial Intelligence, 17:185-203, 1981.

[8] D. G. Lowe. Object recognition from local scale-invariant features. In IEEE International Conference on Computer Vision (ICCV), pages 1150-1157, Kerkyra, Greece, 1999.

[9] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. IEEE Transactions

on Pattern Analysis and Machine Intelligence (TPAMI), 24(4):509-522, 2002.

AUTHORS PROFILE

SUGANYA T. received BE degree in computer science and Engineering from Anna University Chennai, Tamil Nadu, India in 2005.She is currently pursuing her M.E in Computer Science and Engineering from SNS College of engineering, Coimbatore . She has published 3 papers in reputed international and national level conferences Her research interest include Pattern Analysis, Artificial lntelligence, Data Structures

Prof. B. Anuradha obtained her bachelor's degree in Computer Hardware and Software Engineering from Avinashilingam University and Masters Degree in Embedded systems from Anna University and currently pursuing her Ph.D from Anna University of Technology, Coimbatore. She has more than 10 years of teaching experience and currently, she is working as Associate Professor in Department of Computer Science and Engineering, in SNS College of Engineering, Coimbatore, Tamil Nadu. Her areas of interest include Embedded System, Operating Systems and Computer Architecture. She has published 11 papers in reputed international, national level conferences and International journals.

Prof. B. Chellaprabha completed her B.Sc degree from Bharathidasan University in 1993, Master of Computer Applications from Bharathidasan University in 1996 and M.E in Computer Science and Engineering Anna University Chennai in 2007. Presently she is pursing her doctoral degree in Anna University of Technology, Coimbatore. She has more than 15 years of teaching experience. She is working as a Professor and Head of the Department of Computer Science and Engineering, SNS College of Engineering, Coimbatore, Tamil Nadu. Her areas of interest include Wireless sensor networks, Routing protocols, Congestion control and detection and Network Security.