Flickr Distance: A Motion Prediction Approach for Visual Concepts

Prof. B. Anuradha
Department of Computer Science &Engineering, SNS College of Engineering
Coimbatore, India
anu_thambi @rediffmail.com

T.Suganya
Department of Computer Science & Engineering, SNS College of Engineering
Coimbatore, India
sugan_nec@yahoo.com.

Abstract - Image alignment has been studied in different areas of computer vision for hundreds of years, aligning images depicting different scenes remains a challenging problem. Variant to optical flow where an image is aligned to its temporally adjacent frame we propose SIFT flow [1], a method to align an image to its nearest neighbors in a large database containing huge number of scenes. The SIFT flow algorithm consists of matching densely sampled pixel-wise SIFT features between two images. The SIFT features allow unambiguous matching across different scene appearances. The proposed approach combines the concept of Flickr Distance along with the SIFT flow method 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 predicting the motion of images. Here the images are taken from the database and provided as nearest neighbors to a query image. This skeleton can be used in the applications such as motion field prediction from a single image, motion prediction 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.

I. INTRODUCTION

Similarity measurement has been studied for decades and it remains a hot research topic particularly in multimedia literature. There are quite a few vital applications related to this, including indexing, positioning, clustering and annotation. Similarity measurement leads to the Conceptual correlation measurement which in turn calculates the semantic distance between the two image concepts and determines its Flickr Distance [5].

Image alignment, registration and correspondence between images are major topics in computer vision. There are several levels of scenarios in which image alignment depends. The easiest stage, aligning different frames of the same scene, has been studied for the purpose of image stitching and stereo matching. The image alignment problem becomes more challenging for dynamic frames in video clippings, *e.g.* optical flow estimation [7]. The similarities between two contiguous frames in a video are often formulated as an inference of a 2D flow field. In this work, we are paying attention in a new, higher level of image arrangement: aligning two images from dissimilar 3D scenes but sharing similar scene characteristics. Image alignment at the frame level is thus called scene association. 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 levels of objects of the same class, and some objects present in one image might be mislaid in the other. Due to these issues the scene association problem is extremely exigent. Motivated by optical flow methods, which are able to produce dense, pixel-to-pixel association between two images, we propose SIFT flow [1] for adopting the computational framework of optical flow, but by matching SIFT features instead of raw pixels. In SIFT flow, a SIFT features [8] are carved out from each pixel to exemplify local image structures and predetermine contextual information. A distinct flow assessment algorithm is used to match the SIFT features between two images. The use of SIFT features allows vigorous matching across different frame/object appearances and the spatial replica allows matching of objects located at different parts of the frame. Moreover a coarse-to-fine matching scheme is considered to drastically step up the flow estimation process [1].

II. RELATED WORK

Image alignment, image listing or correspondence, is a wide topic in computer vision, computer graphics and medical imaging, casing motion analysis, video compression, shape listing, and object recognition. It is ahead of the reach of this paper to give a detailed review on image alignment. In this section, we will review the image alignment journalism focusing on

(a) *What* to align and the features that is dependable across images, *e.g.* pixels, edges, descriptors;

(b) *Which* way to align, or the demonstration of the alignment, *e.g.* sparse vs. dense, parametric vs. nonparametric;

(c) *How* to align, or the computational aspects to obtain alignment attributes.

In addition, association can be established between two images, or between an image and image models. In image alignment we must first identify the features based on which image association will be established, then find out the image measurement that does not change from one image to another. In stereo and optical flow the brightness fidelity hypothesis was often made for building the association between two images. But researchers came to comprehend that pixel values are not consistent for image matching due to changes of lighting point of view and noise. Features such as phase, filter banks, communal information and slope are used to match images since they are more reliable than pixel values across frames, but they still fall short to deal with radical 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 report for stronger appearance changes, and are confirmed to be effectual in a variety of applications such as visual tracking , optical flow evaluation and object recognition.

The depiction of the association is another important facet of image alignment. One can exploit the information of every pixel to obtain a intense association, or purely use sparse feature points. The form of the association can be pixel-wise dislodgment such as a 1-D disparity map (stereo) and a 2-D flow field (optical flow), or parametric models such as affine and homographic. Although a parametric model can be anticipated from matching every pixel and an intense association can be interpolated from sparse matching classically, pixel-wise displacement is obtained through pixel-wise association, and parametric motion is anticipated from sparse, interest point detection and corresponding. In between the sparse and intense representation, is the association of contours, which has been used in tracking objects and analyzing activity for texture less objects. The fact that the underlying association between scenes is complex and blurred, and detecting contours from frames can be changeable, leads us to seek for intense, pixel-wise association for scene alignment. Estimating intense association between two images is a nontrivial crisis with spatial regularity, *i.e.* the displacements of neighboring pixels have a propensity to be similar.

When the feature values of the two images are close and temporally even, this disarticulation can be formulated as an unremitting variable and the evaluation problem is often condensed to solving PDE's using Euler-Lagrange. When the feature values are different, or other information such as occlusion needs to be taken into description, one can use belief proliferation 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 presumption permit us to solve intense frame matching problems successfully. Image representations, such as color histograms, surface models, segmented regions GIST features, bag of words and spatial pyramids have been anticipated to find similar images at a global level. General to all these representations is short of significant association across dissimilar image regions, and for that reason, spatial structural information of images [3] tends to be unnoticed. Our curiosity is to create intense association between images across scenes, an alignment problem that can be more demanding than aligning images from the same frame and aligning images of the same object class. Our work relates to the chore of co-segmentation [2] that attempt to concurrently segment the common parts of an image pair, and to the problem of shape matching [9] that was used in the circumstance of object recognition.

Stimulated by the modern advances in image alignment and scene parsing, we propose SIFT flow to establish the association between images across frames. In this paper, we will explore the Flickr Distance and SIFT flow algorithm in more depth and will reveal a wide array of applications for SIFT flow.

III. OUR CONTRIBUTION

A set of images for a concept are taken from a source using object identification method. Each concept may consist of different views and progression for images. Effectual image alignment based on the object and its progression pattern, is a tortuous one hence the distance between the two objects is calculated first and based on that, estimation 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 vigorously identifying and aligning complex scene pairs containing momentous spatial difference. An alignment based bulky image database for image analysis and blend is constructed, where

image information is shifted from the nearest neighbors to an inquiry image according to the distance. The proposed work will be helpful in the areas of motion field prediction, pattern analysis and pattern synthesis from a single still image, image listing and object recognition.

IV. FLICKR DISTANCE

As mentioned before, FD is based on representing a concept by constructing an arithmetic model from a set of linked images, and the concept distance is defined by the distance of the two corresponding models. The framework of manipulating 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, is the shrewd arithmetic model that can detain the visual relationships of the concepts well.

A. Visual Concept Pool

To replicate the harmony of concepts in human cognition, the calculation of conceptual correlation should be performed in daily life surroundings. To attain this, we try to pit the arithmetical semantic relations between concepts from a large collection of the daily life photos. To attain a less biased estimation, the image collection should be very large and the source of the images should be autonomous. Fortunately, the online photo sharing website Flickr meet up both circumstances. There are more than 10^9 photos on Flickr, and these photos are uploaded by independent users. 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 dimensions, such as the Jensen-Shannon divergence

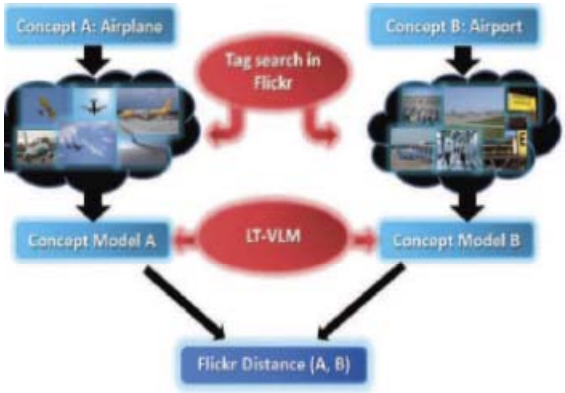

Fig. 1 Determination of Flickr distance.

B. Visual Model assortment

To scrutinize 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 renowned 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.

C. Concept Modeling

Here we intricate on the concept modeling process. A single visual word does not correspond to the specific semantic meaning due to the semantic gap crisis. That is one of the open problems in computer vision. We presume the spatial relationship between the words contains some information. We use the language model, which models the trigrams of the visual words. One diagram may have manifold meanings, while the gist of a trigram has an elevated probability of being exclusive.

V. FEATURE MINING

A. The Sift Flow Algorithm

SIFT is a local feature transform to typify local slope information [8]. In [8], SIFT descriptor is a meager feature representation that consists of both feature extraction and detection. In this paper, however, we only use the feature mining component. For every pixel in an image, we divide its neighborhood into a 4x4 cell array, quantize the orientation into 8 bits in each cell, and obtain a 4x4x8 = 128D vector as the SIFT representation for a pixel. We call this per-pixel SIFT descriptor SIFT image.

To envisage 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 outcrop 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 hallucination in SIFT flow; the whole 128D are used for matching.

B. Matching Objective

We intended an ideal function similar to that of optical flow to estimate SIFT flow from two SIFT images. Similar to optical flow, we want SIFT features to be matched along the flow vectors, and the flow field to be smooth, with discontinuities agreeing with object limitations. Based on these two criterion, the ideal function of SIFT flow is formulated as follows. Let p = (*x, y*) be the grid coordinates 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 believe that there are *L* probable states for *u* (p) and *v* (p), respectively. Let s_1 and s_2 be two SIFT images that we want to match. Set ε contains all the spatial neighborhoods. The power function for SIFT flow is defined as:

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

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

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

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 dislocation term in Eqn. 2 confine the flow vectors to be as small as possible when no other information is available. The smoothness term in Eqn. 3 confine the flow vectors of adjacent pixels to be similar. In this ideal function, shortened 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.

C. Neighborhood of SIFT flow

In theory, we can apply optical flow to two random images to estimate an association, but we may not get a momentous association if the two images are from different scene categories. In fact, even when we relate optical flow to two adjacent frames in a video series, we assume intense 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 inquire a large database with the key in. Ideally, if the database is large and dense enough to contain almost every potential image in the world, the nearest neighbors will be close to the inquiry image, sharing alike local structures.

D. Coarse-to-fine matching scheme

Regardless of the speed we directly optimize Eqn (3) using dual layer belief propagation which balances poorly with respect to image dimension. In SIFT flow a pixel in one image can factually match to any pixels in the other illustration. Suppose the image has h^2 pixels, then $L \approx h$, and the time and space complexity of this dual layer BP is $O(h^4)$. To address the performance hitch, 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 2. For simplicity, we use *s* to represent both s_1 and s_2. A pyramid{s (k)} is constructed, where $s^{(1)} = s$ and $s^{(k+1)}$ is smoothed and sampled from $s^{(k)}$. At each pyramid level *k*, let p_k be the coordinate of the pixel to match, c_k be the compensate or centroid of the probing window, and w (p_k) be the best match from BP. At the top pyramid level $s^{(3)}$, the searching window is centered at p_3 ($c_3 = p_3$) with size mxm, where *m* is the height of $s^{(3)}$. The complexity of BP at this level is $O(m^4)$. After BP congregates, the system propagates the optimized flow vector w (p_3) to the next level to be c_2 where the searching window of p_2 is centered. The size of this searching window is fixed to be nxn with *n* = 11. This practice iterates from $s^{(3)}$ to $s^{(1)}$ until the flow vector w (p_1) is estimated. The difficulty of this coarse-to-fine algorithm is $O(h^2 \log h)$. Moreover, we twice η and hold α and *d* as the algorithm move to a higher level of pyramid in the energy minimization.

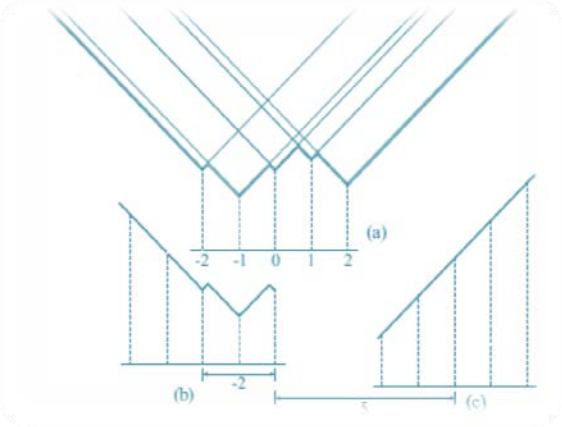

Fig. 2. Distance transform function.

When the matching is propagated from a coarser level to a finer level, the searching windows for two neighboring pixels may have different offsets. We modify the distance transform function developed for condensed L1 norm to manage with this situation, with the idea illustrated in Figure 2. To compute the message passing from pixel p to its neighbor q, we first assemble all other messages and data term, and relate the routine in to the message from p to q. The function is then unmitigated to be outside the range by increasing α per step, as shown in Figure 2 (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 2 (c). If the offset of the searching window for q is –2 otherwise, the message is exposed in Figure 2 (b). Using the proposed coarse-to-fine matching scheme and customized 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 can be achieved through GPU implementation of the BP-S algorithm.

A natural question is whether the coarse-to-fine matching scheme can achieve the same minimum energy as the ordinary matching scheme. 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. 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 3.

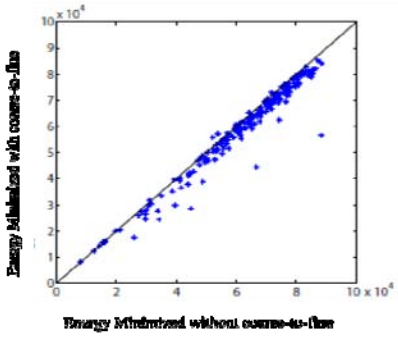

Fig. 3. Coarse-to-fine matching

VI. IMPLEMENTATION

The proposed work is implemented using the modules such as Image analysis, Frames Extraction, All neighbors, Distance Analysis and Motion Prediction.

A. Image Analysis

Different images would be analyzed using digital image processing and the results will be stored in a database. The first step is the conversion with a grey level threshold that segments the foreground objects (white) from the background (black). A segmentation step identifies single foreground objects as objects, which are surrounded by background.

B. Frames Extraction

Features are computed for the remaining objects, which characterize the shape of the objects. Geometric features are based on the object pixels, contour features, which are resulting from the edge pixels (Fourier descriptors, curvature), and features based on the skeleton of the regions were computed. Statistical measures from the vectors express the thickness of the segments. With this step the image processing will be full and the features were used for the examination. In the database all information of the measurement and image processing was stored. The positional information and other metadata are stored together with the image file in the database.

C. All Neighbors

During this module all the possible neighbors of the query image will be captured from the database to perform the matching between images. 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. For a query image, we use a fast indexing technique to retrieve its nearest neighbors that will be further aligned using sift flow. As a fast search we use spatial histogram matching of quantized sift features.

D. Distance Analysis

Similarity between two objects is calculated using a distance measure. This module aims to define a reasonable distance to measure the relationship between the concepts. Each concept corresponds to a visual language model which consists of the trigram conditional distributions under different latent topics.

E. Motion Prediction

This module will robustly identify and aligns complex scene pairs containing significant spatial differences. Aligns objects in similar scenes and allows matching of objects located at different parts of the scene.

VII. CONCLUSION

In this paper, we propose the Flickr Distance to measure conceptual distance. The visual trait of the concepts is modeled by a novel latent topic visual language model. J-S divergence between the Latent Topic Visual Language Model can be considered 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 Normalized Google Distance and Tag Concurrence Distance and we introduced the concept of dense scene alignment: to estimate the dense correspondence between images across scenes. We proposed SIFT flow to match significant local image structures with spatial regularities, and conjectured that matching in a large database using SIFT flow leads to semantically meaningful correspondences for scene arrangement. 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 structure is concretely realized in motion prediction from a single image, motion amalgamation via object relocation and face recognition.

REFERENCES

[1] Ce Liu, Member IEEE, Jenny Yuen, Member IEEE, and Antonio Torralba, Member IEEE - "SIFT Flow: Dense Correspondence across Scenes and its Applications", 2001.
[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 - "Beyond Co-occurrence: Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities", 2012.
[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 - "Flickr Distance: A Relationship Measure for Visual Concepts", 2012.
[6] Zhiyu Zhou, Jianxin Zhang, Li Fang, Zhejiang Sci - China - "Object Tracking Based on Dynamic Template and Motion Prediction.",2009.
[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, 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

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 PhD 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, TamilNadu. 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.

Suganya T. received BE degree in computer science and Engineering from Anna University Chennai, TamilNadu, India in 2005.She is currently pursuing her M.E in Computer Science and Engineering from SNS College of engineering, Coimbatore. She has published 5 papers in reputed international and national level conferences. Her research interest include Pattern Analysis, Artificial intelligence, Data Structures.