Chapter V

Similarity Learning for Motion Estimation

Shaohua Kevin Zhou
Siemens Corporate Research Inc., USA

Jie Shao
Google Inc., USA

Bogdan Georgescu
Siemens Corporate Research Inc., USA

Dorin Comaniciu
Siemens Corporate Research Inc., USA

ABSTRACT

Motion estimation necessitates an appropriate choice of similarity function. Because generic similarity functions derived from simple assumptions are insufficient to model complex yet structured appearance variations in motion estimation, the authors propose to learn a discriminative similarity function to match images under varying appearances by casting image matching into a binary classification problem. They use the LogitBoost algorithm to learn the classifier based on an annotated database that exemplifies the structured appearance variations: An image pair in correspondence is positive and an image pair out of correspondence is negative. To leverage the additional distance structure of negatives, they present a location-sensitive cascade training procedure that bootstraps negatives for later stages of the cascade from the regions closer to the positives, which enables viewing a large number of negatives and steering the training process to yield lower training and test errors. The authors apply the learned similarity function to estimating the motion for the endocardial wall of left ventricle in echocardiography and to performing visual tracking. They obtain improved performances when comparing the learned similarity function with conventional ones.
INTRODUCTION

Image Matching and Similarity Function

Image matching is fundamental to various computer vision tasks. In motion estimation, image matching happens along the temporal dimension, e.g., comparing consecutive frames to establish correspondences over time or tracking points of interest. In image registration, image matching happens along the spatial dimension, e.g., comparing two heterogeneous images for establishing spatial correspondences. Image matching is also vital to content-based retrieval, face recognition, and application of the same kind, where comparing testing and training image is needed.

Underlying an image matching process lays an indispensable component of similarity function. A similarity function is a two-input function \( s(I, I') \) that measures how closely the test patch \( I' \) is visually similar to the template patch \( I \). A typical use of similarity function in, say, motion estimation and image registration algorithms is as follows: given two images \( I \) and \( I' \) and a target point \((u,v)\) whose motion vector or spatial correspondence to be estimated, one finds the shift that has the (local) maximum similarity. If the minimum is sought, one can simply negate the similarity function.

\[
(\delta \hat{u}, \delta \hat{v}) = \arg \max_{(u,v) \in W} s(I(u,v), I'(u+\delta u, v+\delta v)),
\]

(1)

where \( I(u,v) \) is a local patch extracted from the image \( I \), centered at \((u,v)\), and \( W \) is the searching window. In motion estimation, the two images \( I \) and \( I' \) are successive frames, e.g., \( I=I_{t-1} \) and \( I'=I_t \); in image registration, the two images \( I \) and \( I' \) are the image pair to be registered. In retrieval and recognition applications, the use of similarity function is as follows:

\[
\hat{n} = \arg \max_{n=1,2,...,N} s(I_n, I').
\]

(2)

where \{\( I_n; n=1,2,...,N \)\} are gallery images stored in the database, and \( I' \) is a query image that is used to sort the database. The principal difference between (1) and (2) lies in the search space where the maximum is found: The search space for the first type of application (e.g., motion estimation and image registration) is a spatial window, and that for the second type of applications (e.g., retrieval and recognition) is on the index of the images in the gallery database.

In this chapter, we concentrate on the specific application of motion estimation. It is obvious that applications like image registration, retrieval and recognition, etc. can be also tackled with a minor modification.

Similarity Function Under Complex Appearance Variations

Matching two images under complex appearance variations is challenging due to lacking an appropriate similarity function. Similarity functions proposed in the literature are mostly generic and inadequate for handling complex appearance variations. A motivating example is illustrated in Fig. 1.

Consider a stress echocardiographic video (stress echo) (Otto, 2004), a series of 2D ultrasound images of the human heart captured after the patient undergoes exercise or takes special medicine. We focus on wall motion analysis to characterize the functionality of the heart. To be specific, we measure the motion of the endocardium (or the inner border) of the left ventricle (LV). As shown in Fig. 1 the