Feature-based image alignment via coupled Hough transforms

Ron Kimmel, Cuiping Zhang, Alexander M. Bronstein, and Michael M. Bronstein

August 18, 2009

Abstract

We investigate the problem of feature based image alignment under vertical and horizontal translation and scaling. The problem of finding the two dimensional affine alignment without rotation can be separated into two coupled simpler problems. The resulting two model fitting problems are coupled through a consensus set of points that fit both models. The proposed solution can be viewed as a coordinate-wise optimization that separates the two dimensional shift and stretch problem into two coupled one-dimensional problems. Each one dimensional alignment problem is solved through a Hough transform that allows fast and robust estimation of the two parameters of a one dimensional affine deformation model. Experimental results demonstrate the efficiency and accuracy of the proposed method in finding the spatial transform between two given video sequences.

Keywords: video analysis, spatial alignment, scaling, translation, Hough transform.
1 Introduction

In this paper we try to spatially scale and align two given videos based on the given coordinates in space and time and the identities of feature points extracted from some key frames and projected onto a given dictionary. Motivated by human perception and efficient computation, image feature descriptors were proven to be powerful in object matching and recognition in computer vision [11]. One such popular descriptor is the shift invariant feature transform (SIFT) introduced by Lowe [7]. It locates points of interest as extremal values of the Gaussian curvature modulated by the gradient magnitude of smoothed version of the image (aka linear scale space) and captures the local behavior of the image near these points as the local histogram of the gradients, quantized into a small number of bins. An accelerated version is the Bay et al. speeded up robust features (SURF) [1], see alternative approaches [8, 3, 2] and performance evaluation for object retrieval [9, 10, 5].

Here we use the identity (by projection onto a dictionary) and spatial location of such descriptors in order to find the affine minus rotation transformation between two video streams, which is an important problem in video analysis. The inputs to our alignment procedure are two streams of the same content where the videos could have gone through different manipulations and transformations like trans-coding, format translation, or studio editing. In order to robustly solve the video alignment problem, we exploit the intrinsic separability between the horizontal and vertical stretch and shift. For model parameter estimation we apply 2D polar Hough transforms for extracting the deformation of each axis.

The structure of the paper is as follows: In Section 2 we formulate the problem of image alignment based on feature identities and locations. Next, in Section 3 we present our solution. Unlike the common usage of robust solvers for model parameters that are often employed in the context of feature based matching and alignment, here, we use a somewhat simpler transformation that allows us to split the problem and obtain the desired solution from two Hough transformed planes of the matched feature points. Section 4 explores ways to define the overlapping regions as bounding boxes between the aligned images. We also briefly discuss the larger picture of incorporating the proposed procedure as part of a matching two video streams. Section 5 provides some examples and implementation considerations, and Section 6 concludes the paper.

2 Problem Formulation

Consider a gray level image $I : \Omega \subset \mathbb{R}^2 \rightarrow [0, 1]$ to be a scalar smooth function, where $I = 0$ corresponds to black and 1 to white. Besides $I$ we are also given
the image $J(x) = I(Ax + b) + N(x)$, where $x = (x,y)$ denotes the image coordinates, and

$$A = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

where $a_{11}, a_{22}, b_1,$ and $b_2$ are the scalars, or model parameters, we would like to find, and $N(x)$ represents noise in its most general form. Our goal is to extract $A$ and $b$ from a finite small number of corresponding features in both images.

Denote the feature points at image $I$ as the set of $n$ triplets $\{x_i, y_i, f_i\}_{i=1,...,n}$, where $\{x_i, y_i\}$ are the $x,y$-coordinates of the feature, and $f_i$ its identity. Similarly, the $m$ features of image $J$ are denoted by $\{\tilde{x}_i, \tilde{y}_i, \tilde{f}_i\}_{i=1,...,m}$. It is obvious that given a perfect correspondence, two features points could have been sufficient in order to determine the alignment, assuming the line between the two points we use is not parallel to the $x$ or the $y$ coordinates. The problem becomes more interesting as noise is introduced, in terms of both the location and identity of the feature points. Such noise can be the result of cropping or tras-coding of the video at various bit rates and formats or visual information that was added to or partially occlude the video like subtitles, logos, banners or captions, see e.g. Figure 1. Other sources of noise could be the outcome of gamma-correction or other intensity mappings.

An efficient solution to this problem could be useful for other image analysis tasks like object tracking, in which we search for part of the current frame in sequential frames, again based on feature identities and locations. In this case one should ignore features that are the result of interaction between the tracked object and its background in order to produce meaningful results. Luckily, these kind of features are usually inconsistent in time, which could make the proposed model attractive for tracking problems.

### 3 Proposed solution

By close inspection of our problem we notice that one could separate the parameter fitting into two sub-problems; the horizontal alignment with model parameters $a_{11}$ and $b_1$, and the vertical alignment with $a_{22}$ and $b_2$. Let us first explore the problem of finding the parameters for a one dimensional affine transformation

$$x = a\tilde{x} + b,$$

in which we would like to robustly solve for $a$ and $b$ given a set of possible matching couples $\{x_i, \tilde{x}_j\}$ extracted by matching the identities of the feature descriptors. In our problem we may limit the slope $a$ to be close to one using the prior knowledge that a video stretch can be up to $\sim 30 - 40\%$. It is simple to see that we are
Figure 1: Schematic sketch of an image after nonuniform scale and insertion of a subtitle with some corresponding feature points identified in both instances. Our goal is to find the transformation between the two images by matching the feature points (location and identity).
in fact searching for the line defined by \( a \) and \( b \) in the \( \{x, \tilde{x}\} \) plane. Lines can be robustly detected using the Hough transform \([6]\), were specifically we use a polar version of the transform \([4]\) in which we limit the slope. The set of lines passing through a point \( \{x_0, \tilde{x}_0\} \) can be written as \( x_0 \cos \theta + \tilde{x}_0 \sin \theta = \rho \) in the so-called Hough plane, see Figure 2 for a sketch of the polar Hough transform.

![Figure 2: Lines that pass through a point in the \( \{x, \tilde{x}\} \)-coordinates plane (left), are transformed into a sinusoidal curve by the polar (generalized) Hough transform (right).](image)

This simple solution works for both the \( x \) and the \( y \) coordinates. First, we find all features with matching identities in the two images and generate their corresponding matching points in the \( \{x, \tilde{x}\} \) plane from which, by Hough transform, we robustly extract the model parameters \( \{a_{11}, b_1\} \). The same procedure applies to \( \{y, \tilde{y}\} \) and \( \{a_{22}, b_2\} \). Note that in case of severe noise, we could further force the matched features to be the same in both coordinates and thereby couple between the two estimation problems more tightly, see Figure 3.

### 4 Bounding boxes in time sequences

Given a match, we could validate the matched region by corresponding regions that are covered by the matched features. This allows us to define a bounding box of a consensus region that can be important for remote clicking interfaces. The matched regions are defined by corresponding features that cover only regions in which there are corresponding features. If the transformations between a given video and a reference one are similar along the two videos, then, we could assume that the corresponding regions are a union of the corresponding regions for
Figure 3: The two sets of images with features and corresponding bounding boxes demonstrate the effect of accurate and robust bounding box extraction. This is done by forcing the matched features to be the same for the $x$ and $y$ coordinates. The red boxes represent decoupled (wrong) reconstructions while the green bounding boxes are the (right) result of enforcing the same consensus set of features for both $x$ and $y$ coordinates.
each frame. Note that the proposed method is robust to disturbances like banners, subtitles, logos, and trademarks that could appear at different locations.

In some cases we have the liberty to choose the frames to compare. The problem here is to define a representative image in which the features are robust, stable and well spread along the image boundaries, so that we have a good indication of the actual regions we match. I.e. two robust features at two ends of a diagonal are like four features each located at the four edges of the matched bounding box.

Considering the video as a whole, the input to our problem are two sets of \( k \) corresponding frames \( I = \{I_i\}_{i=1}^k \) (taken from the reference video) and \( J = \{J_i\}_{i=1}^k \) (taken from a different video stream), where \( \{I_i, J_i\} \) are time corresponding frames.

The output we expect to produce is,

- One to one spatial correspondence between \( I \) and \( J \). Specifically, for each temporal corresponding frames find the correspondence \( T_i : I_i \rightarrow J_i \).

- The corresponding sub-image matched domains in \( I \) and \( J \) provided as bounding boxes.

- Indication of the level of confidence in the correspondence.

- Indication if the transformations \( T_i \) is stable in time, for example, if \( |T_i - T_j| < \epsilon \) for all \( i, j \).

## 5 Examples

Figure 4 presents a transposed version of the original image \( I \) at the bottom right, and the distorted image \( J \) top left with the detected feature points. The \( \{x, \tilde{x}\} \)-plane of corresponding features appear at the bottom left. The distortion parameters obtained through hough transform that determines a line in the coordinates \( \{x, \tilde{x}\} \)-plane. The extreme matched features along the detected parameters line determine the box bounding all corresponding parameters between the two images.

Next, Figure 5 describes the coordinates planes, \( \{x, \tilde{x}\} \) on the left and \( \{y, \tilde{y}\} \) on the right, and the corresponding polar Hough transforms at the bottom.

That is, a sinusoidal curve in the Hough plane represents all lines that can pass through a given point in the \( \{x, \tilde{x}\} \)-plane. The detected dominant \( \{\rho, \theta\} \)-coordinate in the Hough transform corresponds to the line detected at the coordinate plane.
Figure 4: Top left is the image $J$ with black banners and different aspect ratio than the original image $I$ displayed transposed at the bottom right. Aligning the images and determining the matched bounding boxes is performed through matched feature points at the corresponding \{x, \tilde{x}\}, bottom left, and \{\tilde{y}, y\}, top right, planes.
Figure 5: Top left $\{x, \tilde{x}\}$-coordinates plane, in which each point corresponds to a possible match between features points from $I$ and $J$. Bottom left is the polar hough transform in which all lines that pass through a point in the $x\tilde{x}$ plane correspond to a sinus at the Hough $\{\rho, \theta\}$-plane. The intersection of the largest number of sinus waves indicates consensus for a possible line at the $\{x, \tilde{x}\}$-plane. Right, same for $\{y, \tilde{y}\}$. 
6 Conclusions

We have shown how to decouple the affine minus rotation problem of aligning two images or video streams into two one dimensional affine alignment problems. The one dimensional problems are coupled through a consensus set of corresponding features and are solved by a robust parameter extraction method. For estimating the scaling and translation we use the 2D-Hough transform. Other robust fitting methods are possible. In the future we plan to investigate ways to refine the temporal alignment using the image spatial alignment tool and apply it for tracking objects in videos. We expect the tracking procedure to work as long as there are no substantial rotations.

References
