Learning geometrical transforms between multi camera views using Canonical Correlation Analysis

Christian Conrad
conrad@vsi.cs.uni-frankfurt.de
Rudolf Mester
mester@vsi.cs.uni-frankfurt.de

Visual Sensorics and Information Processing Lab
Goethe University, Frankfurt am Main, Germany
Computer Vision Laboratory
Electr. Eng. Dept. (ISY), Linköping University, Sweden

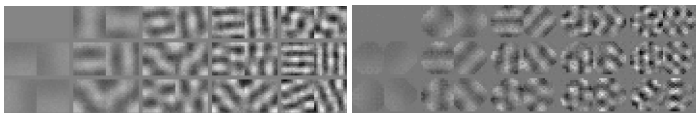

Figure 1: **CCA on natural images:** (left) Filters learnt on pairs of natural images related by a (left) 90 degree rotation and (right) 45 degree rotation.

We present an unsupervised and sampling-free approach to learn the correspondence relations between pairs of cameras in closed form employing a linear model known as Canonical Correlation Analysis (*CCA*). The only assumption we make is that the relative orientation between the cameras involved is fixed. In a two stage algorithm, we first learn the inter-image transformation based on CCA. This analysis usually has to be done in a multi-scale framework, as applying CCA directly to full resolution images may be computationally prohibitive. In the second stage we employ the learnt transformation which is given only implicitly and predict for a given pixel in a first view its corresponding region within a second view. We denote these regions as correspondence prior.

CCA has been introduced by Hotelling [2] as a method of analyzing the relations between two sets of variates and can be applied in closed form. Consider two random vectors $\mathbf{x}$ and $\mathbf{y}$ where $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^M$. The goal of CCA is to find basis vectors for which the correlation between $\mathbf{x}$ and $\mathbf{y}$ when projected onto the basis vectors are mutually maximized [3]. In the case of a single pair of basis vectors $\mathbf{u} \in \mathbb{R}^N, \mathbf{v} \in \mathbb{R}^M$ the projections are given as $a = \mathbf{u}^T\mathbf{x}$ and $b = \mathbf{v}^T\mathbf{y}$. Assuming $\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{y}] = 0$, the correlation ρ between a and b can be written as:

$$\rho(a,b) = \frac{\mathbb{E}[ab]}{\sqrt{\mathbb{E}[aa]\mathbb{E}[bb]}} = \frac{\mathbb{E}[(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})]}{\sqrt{\mathbb{E}[(\mathbf{u}^T\mathbf{x})(\mathbf{u}^T\mathbf{x})]\mathbb{E}[(\mathbf{v}^T\mathbf{y})(\mathbf{v}^T\mathbf{y})]}}, \quad (1)$$

$$= \frac{\mathbf{u}^T\mathbb{E}[\mathbf{x}\mathbf{y}^T]\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbb{E}[\mathbf{x}\mathbf{x}^T]\mathbf{u}\mathbf{v}^T\mathbb{E}[\mathbf{y}\mathbf{y}^T]\mathbf{v}}} = \frac{\mathbf{u}^T\mathbf{C}_{xy}\,\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbf{C}_{xx}\,\mathbf{u}\mathbf{v}^T\mathbf{C}_{yy}\,\mathbf{v}}}. \quad (2)$$

Note that (2) does not depend on the actual scaling of $\mathbf{u}$ or $\mathbf{v}$, therefore in the case of a single pair of basis vectors CCA can formally be defined as solving the following optimization problem:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{C}_{xy}\,\mathbf{v} \qquad \text{s.t.} \quad \mathbf{u}^T\mathbf{C}_{xx}\,\mathbf{u} = \mathbf{v}^T\mathbf{C}_{yy}\,\mathbf{v} = 1. \quad (3)$$

It can be shown that (3) can be cast as a generalized eigenproblem with $K = min(N,M)$ solutions, defining two sets of basis vectors $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ with $k = 1,..,K$ [1]. The projections $a_k = \mathbf{u}_k^T\mathbf{x}$ and $b_k = \mathbf{v}_k^T\mathbf{y}$ are uncorrelated which implies that $\mathbf{u}_i^T\mathbf{C}_{xx}\mathbf{u}_j = 0$ and $\mathbf{v}_i^T\mathbf{C}_{yy}\mathbf{v}_j = 0$ for $i \neq j$. Assuming that $N = M$, and arranging the basis vectors column wise such that $\mathbf{U} = \{\mathbf{u}_k\}$ and $\mathbf{V} = \{\mathbf{v}_k\}$ one can show that $\mathbf{U}$, and $\mathbf{V}$ define a basis for the random vectors $\mathbf{x}$ and $\mathbf{y}$, respectively [1]. Figure 1 shows filters learned with CCA on pairs of natural images.

Given a binocular image stream, we learn the inter-image transformation by applying CCA on data matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{T\times N}$, where each row in $\mathbf{X}$ and $\mathbf{Y}$ correspond to a subsampled version of the original image, respectively. We whiten the two data matrices before applying CCA such that $\mathbf{C}_{xx} = \mathbf{C}_{yy} = \mathbb{I}$ holds. Then the two constraints within (3) relax to $\mathbf{u}^T\mathbf{u} = \mathbf{v}^T\mathbf{v} = 1$ and the sets of basis vectors determined by CCA will form orthonormal bases which allows us to perform the prediction in closed form. We whiten the data matrices using PCA from which we obtain matrices $\mathbf{W}_\mathbf{x}$, $\mathbf{W}_\mathbf{y} \in \mathbb{R}^{N\times N}$ containing the PCA basis vectors for our given data in $\mathbf{X}$, and $\mathbf{Y}$. Given that CCA is able to perfectly learn the transformation then the correlation between a pair of corresponding patches $(\mathbf{x},\mathbf{y})$ will be maximum iff

$$\mathbf{U}_w^T(\mathbf{W}_\mathbf{x}^T\mathbf{x}) = \mathbf{V}_w^T(\mathbf{W}_\mathbf{y}^T\mathbf{y}). \quad (4)$$

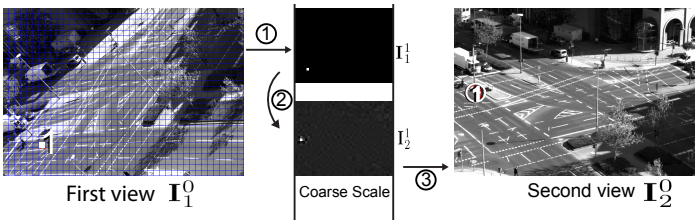

Figure 2: Visual description of how correspondence priors are generated, once the transformation between two views has been learnt via CCA.

This implies that we can predict $\mathbf{y}$ from $\mathbf{x}$ and vice versa which is the same as applying the learnt transformation to $\mathbf{x}$ or $\mathbf{y}$. Solving (4) for $\mathbf{x}$ we can predict $\mathbf{x}$ from $\mathbf{y}$ as:

$$\mathbf{x} = \mathbf{W}_\mathbf{x}(\mathbf{U}_w(\mathbf{V}_w^T(\mathbf{W}_\mathbf{y}^T\mathbf{y}))). \quad (5)$$

Correspondence priors are then generated as follows: Let (x,y) be the spatial coordinates of a pixel in $\mathbf{I}_1^0$, where the superscript 0 denotes the original resolution. Next, determine the pixels' coordinates (u,v) within the low resolution, and generate a binary image of the same size as the low resolution where the pixel at (u,v) is set to 1 (step 1 in Fig. 2). Using (5) we apply the learned transformation to the binary image and obtain the predicted image. When there is a one-to-one pixel correspondence and the transformation has perfectly been determined, the predicted image will be binary again where a single pixel is set to 1 marking the correspondence. However, due to noise one typically obtains predictions that encode regions of high probability containing the correspondence (step 2 in Fig. 2). Interpreting the predicted images as an empirical bivariate correspondence distribution over the spatial image coordinates we encode the prediction by means of a 2×2 covariance matrix $\mathbf{C}_p$. Based on an eigenvalue analysis of $\mathbf{C}_p$, we consider that a correspondence exists if both eigenvalues of $\mathbf{C}_p$ are small. Finally, the correspondence prior to the pixel at $(x,y) \in \mathbf{I}_1^0$ in the second view is given as the covariance error ellipse from $\mathbf{C}_p$ projected onto $\mathbf{I}_2^0$ (step 3 in Fig. 2). Figure 3 shows several correspondence priors and correspondence maps for different real world setups.

[1] T.W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, 1958.

[2] H. Hotelling. Relations between two sets of variates. *Biometrika*, 1936.

[3] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis*. Academic Press, 1980.

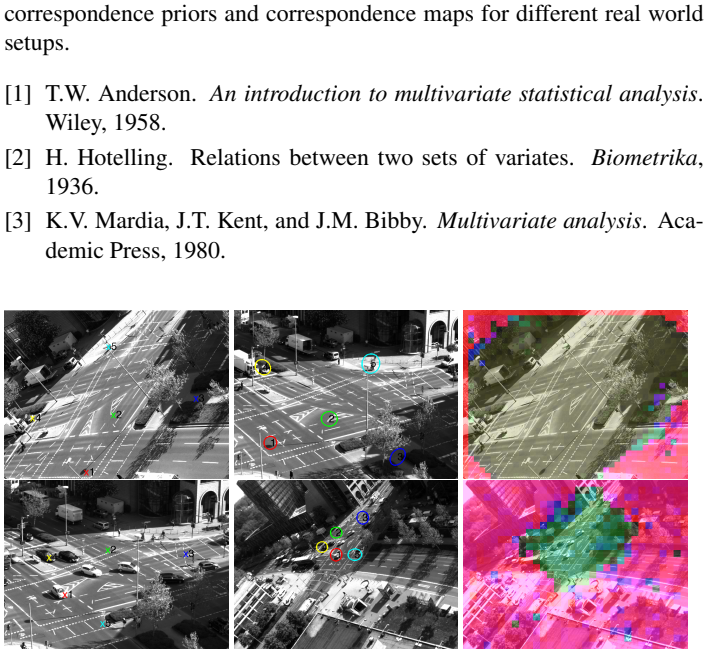

Figure 3: Correspondence priors and maps for real world setups. For selected pixels within the first view (first column), regions of high probability containing the true corresponding pixel in second view (middle column) are determined. (right column) Correspondence maps. Regions colored in shades of purple have a high probability of not being visible in the first view.