DETECTING GESTURES IN MEDIEVAL IMAGES

Joseph Schlecht [1], *Bernd Carqué* [1,2] *and Björn Ommer* [1]

[1] Interdisciplinary Center for Scientific Computing
[2] Institute for European Art History
Ruprecht-Karls-Universität Heidelberg
{schlecht,ommer}@uni-heidelberg.de, b.carque@zegk.uni-heidelberg.de

ABSTRACT

We present a template-based detector for gestures visualized in legal manuscripts of the Middle Ages. Depicted persons possess gestures with specific semantic meaning from the perspective of legal history. The hand drawn gestures exhibit noticeable variation in artistic style, size and orientation. They follow a distinct visual pattern, however, without any perspective effects. We present a method to learn a small set of templates representative of the gesture variability. We apply an efficient version of normalized cross-correlation to vote for gesture position, scale and orientation. Non-parametric kernel density estimation is used to identify hypotheses in voting space, and a discriminative verification step ranks the detections. We demonstrate our method on four types of gestures and show promising detection results.

1. INTRODUCTION

We present an automatic method to find gestures in the illustrations of medieval manuscripts. Our focus on gestures in the visual arts of the Middle Ages is the first step in a long-term interdisciplinary project to gain deeper insight into the nature of embodied communication in medieval culture [15]. We base our approach on four illustrated manuscripts of Eike von Repgow's *Mirror of the Saxons*. The detector described in this paper lays the groundwork to compare corresponding scenes from each copy automatically with regard to the depicted gestures.

Our goal is to detect multiple types of gesture at different scales and orientations in the digitized manuscripts. The fact that the gestures are drawn by hand introduces a significant challenge due to artistic variation. A positive aspect of these man-made images, however, is that they follow simple 2-D patterns without perspective. We take advantage of this drawing style with a template driven detection strategy. Given labeled instances of a particular type of gesture, our approach centers on learning a subset that spans its appearance variation. We cast votes for detections based on an efficient version of normalized cross-correlation, followed by a verification stage to rank the hypotheses.

(a) Heidelberg manuscript excerpt

(b) Enlarged scene from the Heidelberg ms.

(c) Dresden ms.

(d) Wolfenbüttel ms.

Fig. 1. Excerpts from the *Mirror of the Saxons*. (a) is a cropped page from the Heidelberg manuscript; (b) zooms into a scene. (c) and (d) show the same scene in the Dresden and Wolfenbüttel versions.

The high significance of the *Mirror of the Saxons* stems from its outstanding relevance to medieval cultural history. Composed ca. 1220–1235, Eike's text is one of the oldest prose works written in German and, most notably, the earliest German vernacular law book and thus one of the most important monuments in the history of German law [8]. Only four illustrated versions remain, each named after its present location: Heidelberg, Dresden, Wolfenbüttel and Oldenburg [14]. The manuscripts were written, drawn and tinted between ca. 1300 and 1370. A few excerpts can be seen in Fig 1. Along with the illuminated manuscripts of the *Corpus Iuris Civilis* and the *Decretum Gratiani*, they constitute the most famous pictorial heritage of the medieval history of law.

Interaction and communication between persons depicted in the manuscripts is based on characteristic postures of arms, hands or even single fingers. The gestures play a particularly

(a) *pointing* (b) *swearing* (c) *speaking-b* (d) *speaking-f*

Fig. 2. Examples of gestures from the Heidelberg manuscript. Notice that pointing (a) and swearing (b) differ by only a finger. The two speaking gestures are drawn with a back (c) or front (d) view. We demonstrate our detection system on this set of gestures.

important role to researchers of symbolic communication in medieval legal culture [7]. Figure 2 displays a few gestures commonly seen in the manuscripts. To reason about the semantic function of these gestures, it is essential to analyze their historical origin and usage in a detailed, systematic and comparative way. Beyond that, it is necessary to characterize and distinguish the specific handling of gestures by the draftsmen of the different manuscripts. Therefore, art historical analysis will benefit from the comparisons systematically generated by detection algorithms.

Reliable detection of objects in images depends on a good shape representation. Recent works in computer vision represent object shape as a collection of local features [1, 2, 11], parameterized contours [6], and summaries of oriented edges [4]. Some of the more successful approaches learn organized groups of these features [5, 9, 12, 18]. Object templates, on the other hand, offer a dense representation of shape [3, 10, 13, 17], and are particularly effective on objects with a standard configuration, such as faces [16].

Lately, template matching has received less attention. One reason is the availability of sophisticated learning algorithms able to explain the variation of object shape in terms of relationships between local features. With templates, relationships of local structure in an image are fixed. This inhibits the ability to learn generalized parts at the class level. However, templates encode a strong representation of small details that distinguish the object from background and avoid problems of local self-similarity. For our application, we want to detect a repeated and detailed pattern drawn by an artist, e.g., the shape of a hand with fingers in a particular configuration. Feature-based approaches become either intractable or susceptible to clutter and self-similar confusion when representing detailed shape at high-resolution. Thus, we choose to work with an augmented set of templates that accommodate gesture variation. We start with a description of the detection process, then briefly go over how we build the template set and the verification stage.

2. DETECTION

We build our detector from a learned set of templates representing the extent of a gesture's appearance. For each of the templates in this set, we collect their correlation responses to a query image in a Hough accumulator over position, angle and scale. We then search for a set of strong peaks and run a discriminative verification stage that ranks the detections. Later, we compare against using a simpler version of the algorithm with fewer or randomly selected templates to show that understanding the variability of the template is key. We first describe the template detection and voting strategies followed by a brief overview of the verification step.

2.1. Template voting

We begin with set of templates $t_1, \ldots, t_{N_t}$ that capture the appearance variation of a gesture. For a query image f, we apply normalized cross-correlation. The responses are cast as votes into a Hough accumulator over position. The correlation is computed with the mean $\bar{f}_{xy}$ and standard deviation s_{xy} of the image region under a template centered at x, y. The position vote for a template with mean $\bar{t}_i$ is given by

$$h_i(x,y) = \sum_{u,v} \frac{\left[f(x-u, y-v) - \bar{f}_{xy}\right]\left[t_i(u,v) - \bar{t}_i\right]}{M_t \, s_{xy} \, s_t}, \quad (1)$$

where M_t is the template size and s_t its standard deviation. If we let $t'_i = t_i - \bar{t}_i$, and note that it sums to zero, we obtain

$$h_i(x,y) = \sum_{u,v} \frac{f(x-u, y-v)\, t'_i(u,v)}{M_t \, s_{xy} \, s_t}. \quad (2)$$

The numerator is a standard convolution and can be computed efficiently using the Fourier Transform. The denominator, however, poses more of a challenge.

The template standard deviation s_t can be precomputed but s_{xy} cannot, as it depends on each position x, y

$$s_{xy}^2 = \frac{1}{M_t} \sum_{u,v} \left[f(x-u, y-v) - \overline{f}_{xy}\right]^2. \quad (3)$$

It can, however, be efficiently computed with integral images of f and its square [10]. For each position, we use the integral image to look-up the (squared) sum under the area of the template, denoted f_t and f_t^2. Then we have

$$s_{xy}^2 = f_t^2(x,y) - \left[f_t(x,y)\right]^2. \quad (4)$$

We can now efficiently compute the Hough vote. Moreover, we can leverage the Fourier Transform to include a pyramid of weighted Gaussian blur $g_\sigma^1, \ldots, g_\sigma^{N_g}$ over a large range of σ for little cost. The template vote for a position is then

$$h_i(x,y) = \frac{M_t^{-1}}{s_{xy} \, s_t} \mathcal{F}^{-1}\left\{ \mathcal{F}(f)\, \mathcal{F}(t_i) \sum_{n=1}^{N_g} w_n \, \mathcal{F}(g_\sigma^n)^2 \right\}. \quad (5)$$

To detect gestures at various orientation and size, we discretize the angle and scale of templates. Specifically, for a given angle ϕ and scale σ, we transform the templates by $\mathbf{T}_\sigma^\phi$, cross-correlate them with (5) and apply a relatively small

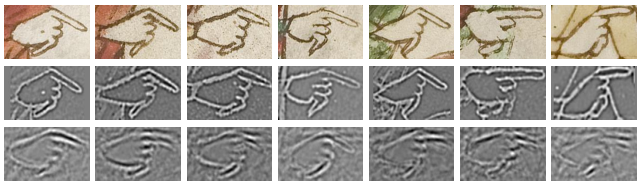

Fig. 3. Subset of pointing gestures with the largest projection on the principle components. The middle row shows their LoG responses to eq. (7); these are the detection templates. The bottom row gives the first seven principle components from left to right.

threshold ρ, summing the result. Thus, we define the full accumulator over gesture position, orientation and scale as

$$H(x, y, \sigma, \phi) = \sum_{i=1}^{N_t} \hat{h}_i\big(x, y\,; \mathbf{T}_\sigma^\phi(t_i)\big)\,. \qquad (6)$$

The thresholded position accumulator $\hat{h}$ reduces the additive effect of small amounts of noise.

We apply a Gaussian kernel density estimator to $H(\cdot)$ to get a non-parametric distribution over voting space. The modes in this density need not be isotropic and can accommodate some details of the gestures we have not parameterized, e.g., aspect ratio. Detection hypotheses are identified by simply searching for local maxima in the discrete density. We keep a number of top scoring hypotheses for each image.

2.2. Hypothesis ranking

The scores from the template detector are normalized with respect to each image, so we apply a discriminative classifier to rank the hypotheses across all images. We compute a histogram of oriented gradients [4] under the oriented bounding box of the hypothesis and evaluate it with a support vector machine. Although this verification scheme works well, evaluating it for even a small number of scales and orientations over all image positions in our high-resolution data set would be prohibitively expensive.

The hypotheses obtained from the template voting stage may not align exactly with a strongly responding HoG feature, so we slide the oriented bounding box a few pixels in the image and re-evaluate, taking the maximum response. The final ranking is a probabilistic estimate based on distance to the support vectors.

3. LEARNING

In this section we briefly describe the template selection process and how we train the verification stage.

3.1. Templates

We extract ground-truth gestures labeled with oriented bounding boxes from a set of training images. We rotate and scale normalize them, and convolve them with the Laplacian of Gaussian (7) to construct our set of templates,

$$\Delta g_\sigma(x, y) = \frac{x^2 + y^2 - 2\sigma^2}{2\pi\sigma^8} \exp\left[-\frac{x^2 + y^2}{2\sigma^2}\right]. \qquad (7)$$

The resulting template pixels have both positive and negative values, and give a strong response when contours closely align. We further scale them to sum to zero. Figure 3 shows a few examples.

Our goal is to find a small collection of templates representative of a particular gesture's appearance. To do this we first compute the principle components $\mathbf{U}$ on the set of gesture templates. We then find the subset with the largest projection on the first N_k principle components using $z_i = \sum_{k=1}^{N_k} t_i^{\mathrm{T}} u_k$. We select the templates with the N_t largest values of z_i.

3.2. Verification

An SVM for verification is trained in two steps. First, we use positive examples of a targeted gesture and negative examples of the other gestures. We then randomly sample detection windows in the background of the training images for negative examples and record false positives. The SVM is retrained with the gestures and background samples for the final detector. We use an RBF kernel and find its parameters using 5-fold cross-validation on the training set.

4. RESULTS AND DISCUSSION

Our ground-truth data set comprises 280 high-resolution scenes from the Heidelberg manuscript. Together, the scenes contain 347 gestures labeled with oriented bounding boxes, each roughly 200×100 pixels. We evaluate the detector with a leave-one-out strategy; we hold out one scene from the manuscript for testing and train on the remaining scenes. We do this for each of the four gesture types in Figure 2. We follow the VOC criteria for correct detections: $A_\cap / A_\cup \geq 0.5$.

The detector is configured as follows. Test images and templates are filtered with $\Delta g_{\sigma=1.5}$. Both N_t and N_k are set to 10. The number of Gaussians N_g is 60 with $g^n_{\sigma\in(0,6]}$. The weights w for each Gaussian are uniform; we have not extensively experimented with other values. For voting, we set $\rho = 40$ and discretize x, y, ϕ, σ-space into $512^2 \times 30 \times 10$ bins. The angle and scale ranges are estimated from training data. Kernel bandwidth is 5 bins over x, y and 1 for ϕ, σ. We limit the number of ranked hypotheses per image to 100.

Precision-recall curves for the detector on four gestures are shown in Figure 5. As a baseline comparison, we evaluated the detector using a single template selected at random against one and five templates with maximum projection on the principle components. Table 1 shows our detection rate and recovery of the ground-truth orientation statistics.

These results show that a small set of templates can capture the primary variation of a gesture and be effective for detection. We further report that about 75% of all the gestures were detected with 1 false positive per image. We plan to continue this work by evaluating the detector on other versions of

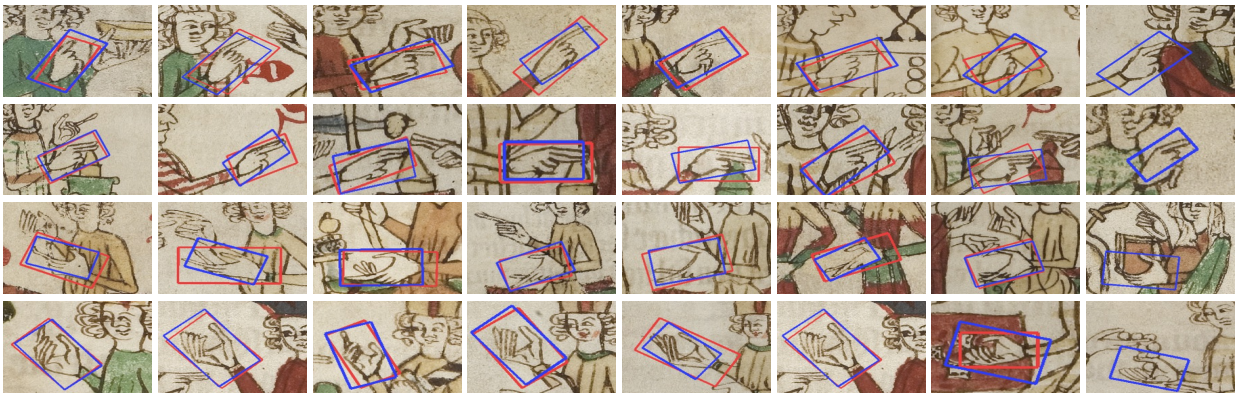

Fig. 4. Example detections of the four gesture types (rows). The ground-truth labels are marked red and the detections are blue. The rightmost column shows a false positive for each gesture. Notice how the false positives for the *speaking* gestures are semantic mistakes.

	N	ϕ	$\pm$	$\widehat{N}$	$\widehat{\phi}$	$\widehat{\pm}$
pointing	98	314.0	17.5	86	317.9	15.7
swearing	99	341.7	14.7	82	344.2	13.3
speaking-b	77	171.7	18.1	67	176.8	15.8
speaking-f	73	226.1	25.0	69	227.7	13.9

Table 1. Gesture frequency N, mean orientation ϕ, and angle standard deviation $\pm$ in the ground-truth data. The detected gesture statistics $(\widehat{\cdot})$ are all very close to the ground-truth. The angles are measured in degrees, clockwise from the x-axis.

the *Mirror of the Saxons* and automating a comparative analysis of legal gestures used in medieval manuscripts.

5. ACKNOWLEDGEMENTS

This work was supported by the Excellence Initiative of the German Federal Government and the Frontier fund.

6. REFERENCES

[1] S. Belongie and J. Malik. Matching shapes. In *ICCV*, pp. 454–461, 2001.

[2] A. C. Berg and J. Malik. Geometric blur for template matching. In *CVPR*, pp. 607–615, 2001.

[3] R. Brunelli. *Template Matching Techniques in Computer Vision: Theory and Practice*. 2009.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pp. 886–893, June 2005.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comp. Vision*, 61(1):55–79, 2005.

[6] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(1):36–51, 2008.

[7] G. Kocher. *Zeichen und Symbole des Rechts*. 1992.

[8] H. Kümper. *Sachsenrecht*. 2009.

[9] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Statistical Learning Workshop*, 2004.

[10] J. P. Lewis. Fast template matching. In *Canadian Image Processing and Pattern Recognition*, pp. 120–123, 1995.

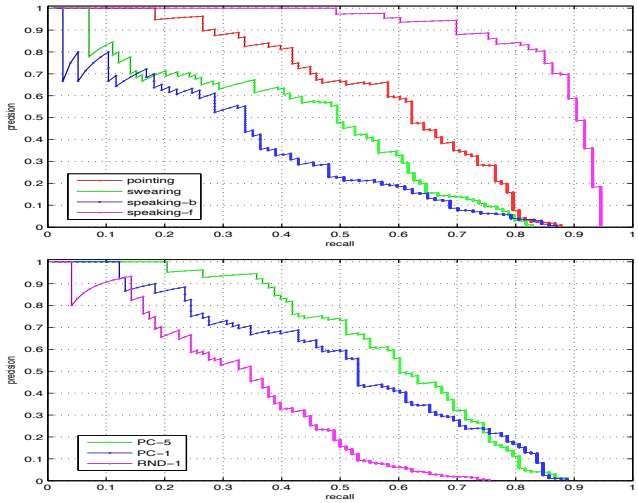

Fig. 5. Precision-recall curves for the presented gesture detection method. (**top**) Detection results over all gesture types using the full system. (**bottom**) Comparison of the *pointing* gesture using 1 (blue) or 5 (green) templates near the principle components against a single template selected at random (magenta).

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comp. Vision*, 60(2):91–110, 2004.

[12] A. Opelt, A. Pinz, and A. Zisserman. Learning an alphabet of shape and appearance for multi-class object detection. *Int. J. Comp. Vision*, 80(1):16–44, 2008.

[13] B. Reddy and B. Chatterji. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Trans. Image Processing*, 5(8):1266–1271, 1996.

[14] R. Schmidt-Wiegand and W. Milde, eds. *Gott ist selber Recht. Die vier Bilderhandschriften des Sachsenspiegels*. 1992.

[15] J.-C. Schmitt. *La raison des gestes dans l'Occident médiéval*. 1990.

[16] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *CVPR*, pp. 586–591, 1991.

[17] G. Wolberg and S. Zokai. Robust image registration using log-polar transform. In *IEEE Int. Conf. Image Processing*, 2000.

[18] P. Yarlagadda, A. Monroy, and B. Ommer. Voting by grouping dependent parts. In *ECCV*, pp. 197–210, 2010.