2009-03-11

An alignment based similarity measure for hand detection in cluttered sign language video

Boston University, Computer Science technical report BUCS-TR-2009-009

An alignment based similarity measure for hand detection in cluttered sign language video

Ashwin Thangali and Stan Sclaroff
Department of Coumputer Science, Boston Universrty
111 Cummington St, Rm 138, Boston, Ma 02215
[tvashwin,sclaroff]@cs.bu.edu, http://www.cs.bu.edu/groups/ivc/

Abstract

Locating hands in sign language video is challenging due to a number of factors. Hand appearance varies widely across signers due to anthropometric variations and varying levels of signer proficiency. Video can be captured under varying illumination, camera resolutions, and levels of scene clutter, e.g., high-res video captured in a studio vs. low-res video gathered by a web cam in a user's home. Moreover, the signers' clothing varies, e.g., skin-toned clothing vs. contrasting clothing, short-sleeved vs. long-sleeved shirts, etc. In this work, the hand detection problem is addressed in an appearance matching framework. The Histogram of Oriented Gradient (HOG) based matching score function is reformulated to allow non-rigid alignment between pairs of images to account for hand shape variation. The resulting alignment score is used within a Support Vector Machine hand/not-hand classifier for hand detection. The new matching score function yields improved performance (in ROC area and hand detection rate) over the Vocabulary Guided Pyramid Match Kernel (VGPMK) and the traditional, rigid HOG distance on American Sign Language video gestured by expert signers. The proposed match score function is computationally less expensive (for training and testing), has fewer parameters and is less sensitive to parameter settings than VGPMK. The proposed detector works well on test sequences from an inexpert signer in a non-studio setting with cluttered background.

1. Introduction

In this paper, we focus on hand detection in American Sign Language (ASL) video sequences captured in both controlled and uncontrolled settings. We envision future systems for ASL gesture recognition and gesture based retrieval that enable users to search through sign language video (videos could be from stories, news media, lectures, performances, reference sources, and instructional material) via gestures to a web cam. As an interim goal, we are developing a query-by-sign ASL lexicon system, where queries are signs gestured by inexpert signers to assist in their learning of sign language. Accurate hand location detection is an essential component for these applications to enable subsequent steps such as hand tracking, hand pose estimation, and hand shape classification.

Linguists have identified approximately 84 distinct hand shapes commonly employed in ASL [21]. Hand shapes oriented in different directions in space can convey distinct signs. Linguistic production constraints reduce the possible range of hand shapes within a single sign and often enforce hand shape symmetry for two handed signs; we have not leveraged these constraints in our current work. The richness and large space of possible hand shapes compounded with factors listed below make hand analysis in sign language video challenging.

- Between signer variations: two signers for the same sign may use slightly (sometimes significantly) different hand shapes and hand orientations, anthropometric and gender differences are typical, the signers may have different ASL proficiencies and learning background.
- Occlusions: hands occlude each other, oftentimes the hand is in front or close to the face causing ambiguity between hand and background.
- Changing environment: background clutter, clothing, illumination, scale and perspective changes are common issues to contend with. Motion blur and image sensor noise are magnified in indoor environments.
- Annotation inaccuracies: in our ASL video sets annotated with hand locations, there is variation in the tightness and centering of bounding boxes. Positioning boxes accurately is difficult to do when hands are close or interacting with each other. The algorithm for hand detection should be robust to these inaccuracies.

Our proposed approach for hand detection reformulates the Histogram of Oriented Gradient (HOG) [6, 16] representation with an explicit alignment step to allow for non-rigid deformations between pairs of image chips[1]. HOG feature descriptors are extracted from overlapping patches

[1] We use the term image chip to denote a sub-image or a region of interest (ROI) within an image.

with centers on a regular feature point grid within an image chip. Traditional HOG based distance measures assume a one-to-one spatial correspondence between the feature points in the two image chips. For instance, Dalal and Triggs [6] employ a linear kernel for pedestrian detection, which corresponds to a dot product between the two sets of HOG features. We propose an aligned distance between a pair of chips computed using the best matching HOG feature vector in the local neighborhood for each HOG feature location.

Approaches robust to large viewpoint changes have been proposed for the object recognition task. The Pyramid Match Kernel (PMK) [11] and Proximity Distribution Kernel (PDK) [15] are two approaches that use feature space partitioning (or discretization) techniques to create a histogram representation of all feature vectors from an image chip. A histogram intersection score gives the similarity measure between a pair of image chips.

We formulate a SVM classifier for hand vs. not-hand image chip classification using the aligned distance. We show improved hand detection accuracy (in terms of ROC area and detection rate) over the rigid match and VGPMK based classifiers in studio sequences of signs signed by expert signers. The detector in this experiment is trained on hand images from a female signer and tested with a male signer. We show that the approach works well on test webcam quality video sequences gestured by an inexpert signer (the intended group for our ASL lexicon application) wearing low skin tone contrast clothing and with background clutter. In this case, the detector is trained with hand images from both expert signers collected in studio.

2. Related work

Several approaches address the problem of hand detection and tracking for general hand gesture recognition: a 2D graphical model for finger articulation is proposed in [24] and approaches using a 3D hand model are proposed in [20, 7]. While there are a wide range of possible hand shapes and poses in sign language, the hand shapes are also highly structured and are generated by linguistic rules. This necessitates approaches specific to the sign language setting.

Farhadi, et al. [9] survey previous work in HMM models for sign recognition and formulate the sign recognition task as a transfer learning problem; given a dictionary of sign videos from one signer and test videos from another signer for a subset of the dictionary gestures, the authors demonstrate that sign classifiers can be automatically built for the remaining set of test gestures. They demonstrate good results on a test signer wearing a long sleeved shirt against a plain background (the dictionary sequences were synthetically generated from a computer graphics avatar). The authors use SIFT descriptors extracted on a regular grid as appearance descriptors for hand images.

Buehler, et al. [4] are motivated by a similar problem to ours: handling variations across singers, cluttered background, two hands often being close or interacting with each other. They address this problem in a tracking setting using a pictorial structure model for the upper body. The authors initialize the model for 5% of frames in the test video. We model hand appearance for hand detection and do not need initialization; hence our approach is complementary to their work.

Ong and Bowden [18] use block difference based features chosen via AdaBoost training for hand detection. These features are suited for high contrast settings (e.g., light colored hands that have good contrast against a dark background and signers wear long sleeved shirts). The problem of aligning hand shapes between signers of different proficiencies is not addressed since weak classifiers chosen during training use difference of image blocks at fixed locations within the ROI

Derpanis, et al. [8] decompose ASL gestures into 14 phonemic movement elements and derive a mapping between these elements and hand trajectories in the image plane. Skin detection and frame-differencing are employed for hand tracking. Motion signatures derived from time series of hand trajectories are mapped to phonemes. The authors demonstrate good phonemic recognition rates using these signatures.

Athitsos and Sclaroff [2] present a method to match lines extracted from synthetic images of ASL hand shapes to edges in real hand images with cluttered backgrounds. The authors demonstrate improved performance over chamfer distance for static hand shapes. The authors in [1] show that chamfer distance is not well suited for ASL hand shape matching. Our application is targeted for hand detection in ASL sequences; motion blur and large between signer variations make it infeasible to match with synthetic hand images.

Hamada, et al. [13] propose a hand contour alignment approach for hand detection and hand pose estimation. They show results with the same signer in training and test sequences wearing long sleeved shirt captured against a simple background.

Yuan, et al. [25] formulate the hand detection problem as a function parameterized by hand shape. This allows the detector for different hand shapes to be trained jointly while allowing a detector tuned to a specific hand shape to be sampled at test time. The authors use a dot product between HOG feature vectors.

To cope with occlusions, Fujimura and Xu [10] propose an algorithm to separate hand blobs when hands are interacting with each other. The authors use depth images to segment the hand regions and propose a skeleton graph partitioning method to separate interacting hands.

Smith, et al. [19] propose a method for resolving hand over face occlusions by modelling background clutter using an image force field. They evaluate their approach on non-sign language gestures.

Hierarchical representations with the bag-of-features model allow for flexible matching between two sets of image features and have shown good performance on object and category recognition tasks. The Pyramid Match Kernel (PMK) [11] and its extension, the Vocabulary Guided PMK (VGPMK) [12] represent a set of HOG features from an image chip as a multi-resolution histogram in the feature space. HOG feature vectors are augmented with (x, y) coordinates of the corresponding feature point to encode spatial proximity information. We use VGPMK in hand detection experiments for comparison with the proposed approach since its performance was shown to be better than PMK for high dimensional features [12].

Ling and Soatto [15] propose a code book representation of feature vectors extracted from training image chips. A histogram is constructed to capture the spatial (x, y) proximity for all pairs (or triples) of code book elements within an image chip. The authors demonstrate performance improvement on object category recognition data sets. The PDK histogram (unlike the PMK representation) only stores pair-wise (or three-wise) proximity information for image features; the global spatial structure of the image features is lost. For hands, we believe it is essential to retain the overall spatial structure within the hand image.

3. Aligned distance measure for image chips

Given a pair of image chips I_1, I_2 normalized to a fixed size, we define an aligned distance score that allows for non-rigid deformations between the images. This is essential for matching hand chips due to the flexibility and variance inherent in hand shape and pose across signers.

Histogram of Oriented Gradient [16, 6] descriptors for image chips (examples illustrated in Figure 1) are extracted as follows. We define a 2-D grid of uniformly spaced feature point locations $\mathcal{G} = \{(x_i, y_j) : i = 1 \ldots G, j = 1 \ldots G\}$ within the image chip. In our implementation, image chips are 90×90 pixels and an 8×8 feature point grid is defined at one scale. Image patches of size 20×20 pixels with centers $\mathcal{G}_{i,j}$ form image regions for HOG feature extraction. Adjacent patches overlap by 10 pixels to allow non-rigid alignment computation between a pair of image chips. The color gradient at a pixel is computed as the maximum magnitude gradient vector in the RGB color planes and a Sobel operator with a Gaussian smoothing filter is used for gradient computation. Each HOG image patch is subdivided into 2×2 cells. Gradient magnitudes within each cell are accumulated into 12 orientation bins over the range $[0, 2\pi)$. Feature vectors from cells in a patch are concatenated to form a 48 dimensional HOG feature vector. This vector is then normalized to unit length for robustness to illumination and contrast changes (as was proposed for the SIFT descriptor in [16]). Thus, we represent a HOG feature vector for an image patch at grid location (x_i, y_j) by $H_{i,j} \in \mathbb{R}^{48}$.

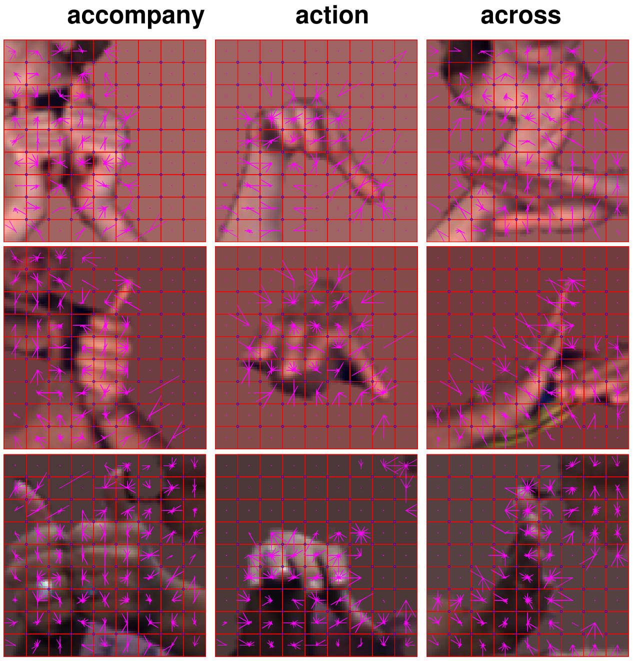

Figure 1. HOG feature extraction for hand image chips from three signers for the same signs. Variation in hand shape across signers necessitates the distance measure between image chips to allow for non-rigid deformation. The blue circles are the centers for HOG patches, each HOG patch corresponds to a 2×2 block of 10×10 pixel cells shown here with red boxes. Adjacent HOG patches overlap by one cell width. A 48 dimensional feature vector normalized to unit length is used for each patch. Even though cells are shared by two or more HOG patches, their contribution to each HOG feature vector is different due to the normalization step.

Let $\mathcal{N}_{i,j}$ be the set of feature locations in the spatial neighborhood of $(x_i, y_j) \in \mathcal{G}$ within distance $T_{\mathcal{N}}$,

$$\mathcal{N}_{i,j} = \{(k, l) : \|(x_i, y_j) - (x_k, y_l)\| \le T_{\mathcal{N}},\ (x_k, y_l) \in \mathcal{G}\}.$$

Our proposed distance function incorporating alignment is given by,

$$D(I_1 \to I_2) = \sum_{\substack{i = 1 \ldots G \\ j = 1 \ldots G}} \min_{(k,l) \in \mathcal{N}_{i,j}} \left\| H^{I_1}_{i,j} - H^{I_2}_{k,l} \right\|. \quad (1)$$

Here, $\| \cdot \|$ is the Euclidean distance between HOG feature vectors. A symmetric distance measure is obtained by adding the directed distance scores,

$$D(I_1, I_2) = D(I_1 \to I_2) + D(I_2 \to I_1). \quad (2)$$

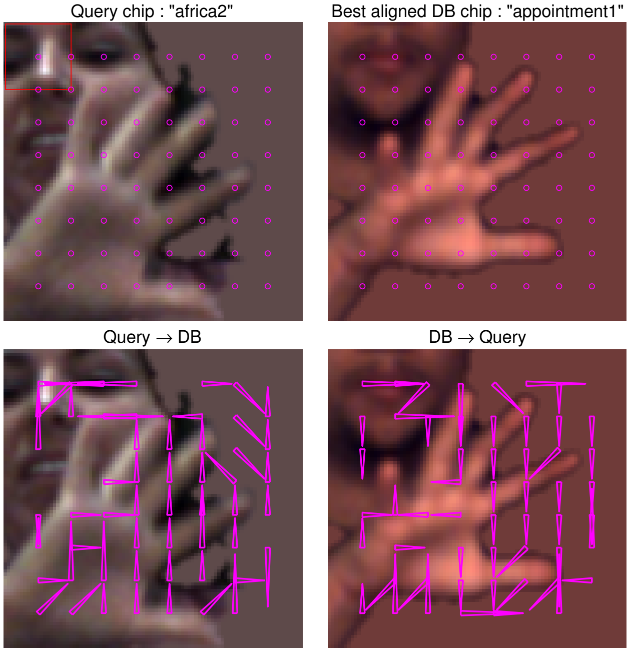

Figure 2. Alignment vectors computed using Equation 1 between a query hand image and the top match database image retrieved using the symmetric distance in Equation 2. The red box in top left image is the ROI for a HOG patch, the circles represent centers of ROIs for HOG patches.

Rigid matching without alignment corresponds to $T_{\mathcal{N}} = 0$. In our implementation, we choose $T_{\mathcal{N}}$ such that $|\mathcal{N}_{i,j}| \leq 13$ (up to 13 neighbors for each feature location, feature points near the image chip boundary have fewer neighbors). We tried $5, 9$ and 13 local neighborhood sizes. While each improves results over rigid match, the 13 neighborhood gave the best results on our data sets.

Aligned distance computation is more expensive than rigid match by a factor of $2 \times average(|\mathcal{N}_{i,j}|)$. The cost in the inner distance computation loop can be reduced by using an early stopping criterion keeping track of the current minimum match score. Larger neighborhood values make the distance computation expensive to run. During hand detection using a scanning window, it is, for instance, possible to reuse some computation from adjacent windows to make the computation more efficient.

4. SVM formulation for hand detection

A distance or similarity measure between pairs of samples is a natural fit for the Support Vector Machine (SVM) formulation. PMK, PDK and Intermediate Matching Kernel [11, 15, 3] are a few examples of approaches that employ this method for object recognition. We use the aligned distance (Equation 2) within a SVM framework for hand/not-hand classification. In our formulation, we consider various hand shapes and hand poses as the foreground class, image chips that partly overlap or do not overlap with hands are considered as the background class. We define the function $K(I_1, I_2)$ for use as the kernel function in a SVM,

$$K(I_1, I_2) = \exp(-\gamma \, D(I_1, I_2)). \tag{3}$$

We note that semi-definiteness of K is not guaranteed. This has been observed by authors in the past with other alignment based distance functions, for instance the chamfer distance and the Hausdorff distance. Grauman, et al. [11] provide an in-depth analysis of various alignment based kernels. In some cases, approximations that satisfy semi-definiteness are possible; for instance, Odone, et al. [17] propose an approximation to the Hausdorff distance and Boughorbel, et al. [3] show an approximation to a version of the aligned distance. In practice, we found using K for hand detection gives stable results, i.e., the quadratic SVM optimization converges to the desired optimum.

A key advantage of our proposed approach in comparison to other kernels like PMK or PDK is that γ is the only parameter to specify. The role of γ is similar to the bandwidth parameter in RBF kernels. The neighborhood size parameter $T_{\mathcal{N}}$ is governed by computational considerations, the expected deformation and image scale. We found larger neighborhood sizes typically work better and in our experiments a 13 neighborhood was used. SVM training converged correctly on our hand detection data sets for various γ values > 0.03 without additional modifications to the kernel matrix to enforce semi-definiteness. Using a 13 neighborhood for alignment, $\gamma < 0.02$ sometimes yields incorrect results for SVM training (the optimized classifier inverts polarity of positive and negative samples). We observed similar behavior with VGPMK for some parameter settings.

5. Hand detection in cluttered ASL video

Our hand detection pipeline follows the standard image scan approach. The sequence of steps is illustrated in Figure 3 for a non-studio sequence from the LAB-F data set. To reduce the computational expense, we use image scan Regions of Interest (ROIs) at one scale and prune the ROIs with the detected skin mask. We run the hand/not-hand SVM classifier for the reduced set of ROIs and choose the top N boxes subject to an overlap constraint for detected hand locations. For all sequences in this set, we used the following parameters for image scan: image scale 82% of original size, ROI dimensions 90 × 90 pixels, ROI spacing 12 pixels, skin mask overlap area for each ROI $> 30\%$, overlapping area between top N detected hand ROIs $< 80\%$.

6. Image pre-processing

We use skin detection to reduce artifacts of clothing changes and background clutter. To achieve additional ro-

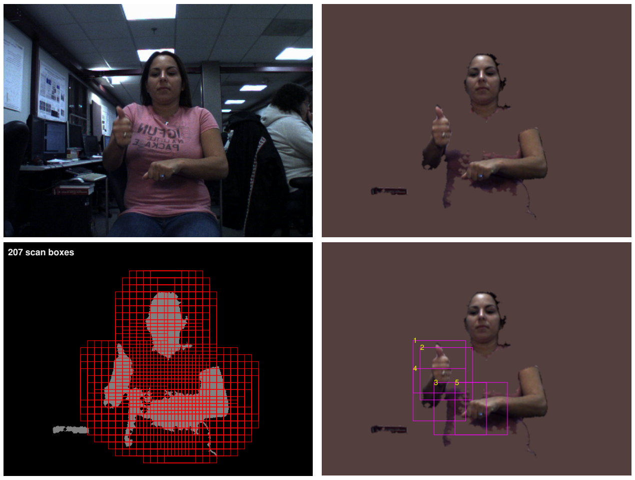

Figure 3. Processing pipeline for hand detection in cluttered ASL video. Results of skin detection and background substitution with average face skin color on an input frame from the LAB-F set are shown in the top row. Image scan ROIs pruned to overlap with the skin mask are shown in the lower left image. HOG features from the pruned ROIs (approximately 200 boxes) are input to the aligned distance based hand detector. The top five detected hand locations are shown in the lower right image. Data used for training the hand detector is described in Section 7.3.

Dataset ID	Signer gender, ASL proficiency, video capture location	# ASL signs	# hand chips
STUDIO-F	female, native, studio	997	64k
STUDIO-M	male, native, studio	680	50k
LAB-F	female, two years, computer lab	605	–

Table 1. Statistics for ASL video sets used in our experiments. The signers sign words from the Gallaudet Dictionary [22]. In the studio setting, we capture 60fps uncompressed videos, with plain background, dark clothing and controlled illumination. These videos have minimal motion blur and good dynamic range. The non-studio video set was captured with a different camera at 30fps compressed in MPEG4 format. All videos are 640 × 480 pixels. Hand location annotations are not available for the LAB-F data set.

bustness to clutter in HOG feature extraction, we substitute background pixels with the average face skin color for all video sequences (this helps since we use color gradients as described in Section 3). We use the Viola Jones detector [23] to detect faces. Histograms in RGB space trained with skin color from STUDIO sequences and background color from a lab background sequence are used to model foreground and background color distributions. A pixel-wise likelihood ratio test is used as the skin color classifier. In frames where a face is not detected, we use average skin segment color to substitute for background pixels. Figures 3, 4 show the results of pre-processing on images from STUDIO and LAB sequences.

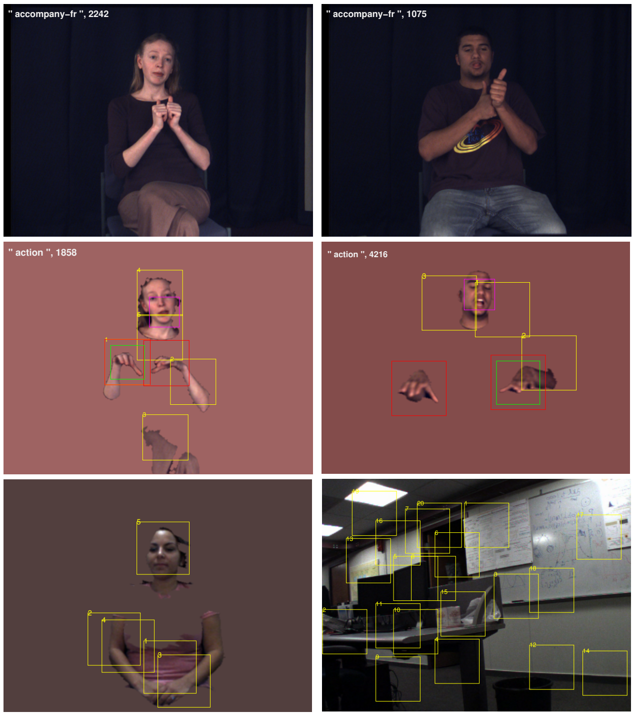

Figure 4. Images in first two rows are from STUDIO-F and STUDIO-M sequences. Results of skin detection based pre-processing are shown in the second row. Magenta boxes are the detected face locations, green boxes are annotated hand locations, red boxes are hand locations resized to ensure uniform scale for hands across the data set. Hand image chips extracted from STUDIO signers are used as foreground samples to train the hand detector. Image chips chosen to overlap $> 20\%$ with detected skin region but overlap $< 60\%$ with hand ROIs are background samples shown here by yellow boxes. Additional background samples are extracted from a rest pose sequence of our test signer and from a lab video sequence. These are shown in the third row.

7. Experiments and performance evaluation

In the first experiment, we compare performance of the aligned distance measure with the rigid match distance and VGPMK [12] for hand detection on STUDIO datasets. In the second experiment, we show hand detection results on video sequences collected with cluttered background from the LAB-F dataset.

7.1. Training and test sets for hand detection

We captured ASL video from three signers; two sets (STUDIO-F and STUDIO-M) were collected from native male and female signers in a photographic studio and one set (LAB-F) was collected from an inexpert signer in the computer lab. The statistics are summarized in Table 1. Our procedure to extract training image chips for hand detection is illustrated in Figure 4.

7.2. Hand detection performance comparison

We use the studio data sets for detector training and testing to quantify improvement in hand detection performance using the aligned distance based detector (hand location annotations are not yet available for the computer lab sequences). To measure generalization performance across signers, we use the STUDIO-F set for training and STUDIO-M set for testing. The training and test sets each contain $4,000$ foreground (hand) and $8,000$ background image chips. Foreground class samples for both hands are sampled from STUDIO-F set in training and STUDIO-M set in testing. Background class samples are extracted from the corresponding foreground sequences, from a test signer sequence and a lab sequence as summarized in Figure 4. All image chips are normalized to 90×90 pixels.

For rigid and aligned distance functions, we use an 8×8 grid of HOG patches as illustrated in Figure 1. In the case of VGPMK, we use a 14×14 grid of HOG patches with other HOG parameters the same as for the rigid distance. A larger set of feature vectors is needed to have sufficient samples to build the VGPMK histogram. Spatial information in VGPMK is encoded by appending the within-image chip feature location $(x, y) \in [0, 1]$ to the normalized HOG feature vector. We used the LIBPMK [14] package to build the VGPMK pyramid and kernel matrices. The bin weights are set as BIN_WEIGHT_INPUT_SPECIFIC and kernel normalization with the diagonal is enabled. VGPMK performance is linked to the parameters used for hierarchical K-means clustering to construct the space partitioning; we tried the following set of parameters, $\{\text{number of levels} \in [5, 7]\} \times \{\text{branching factor} \in [8, 50]\}$. We found that the optimal VGPMK parameters were specific to a data set.

SVM training and test details for the hand detection task are as follows. We use the two-class ν-SVM implementation from the LIBSVM [5] package to train the hand detector. We fix $\nu = 0.005$ in all the experiments. γ in Equation 3 is the only parameter for rigid and aligned distance functions (we use a 13 neighborhood for aligned distance). We sample γ in the range $[0.015, 0.026]$ for the rigid match distance and in the range $[0.04, 0.046]$ for the aligned distance function. The performance of both approaches is not very sensitive to choice of γ (for aligned distance γ should be > 0.03). The results shown in Figure 5 demonstrate that the aligned distance based detector performs better in both ROC area and hand detection rate than rigid match and VGPMK. The best detection rate for VGPMK at 2% false positive rate was obtained with branching factor $= 42$, # levels $= 5$ and yields a detection rate of 94.1% and ROC area $= 0.9937$.

The training and test times for the three algorithms are shown in Table 2. VGPMK needs $\approx$ 5.6Gb of memory for training and testing compared to $<$ 2Gb for rigid and aligned match detectors.

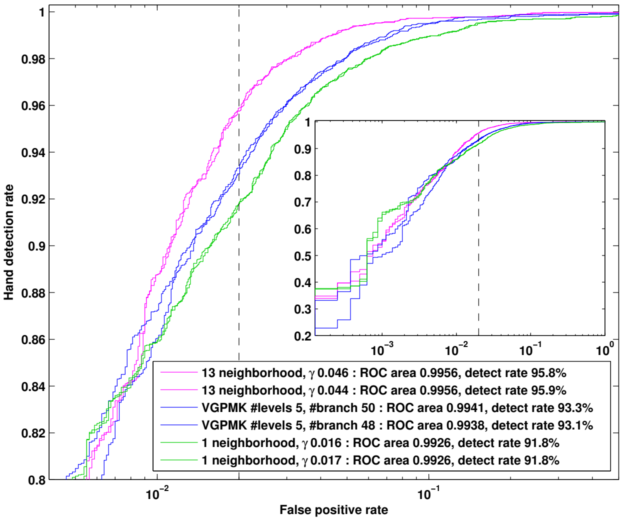

Figure 5. Hand detection ROC curves comparing performance of the aligned distance based classifier with rigid match distance kernel and VGPMK. Training image chips for this experiment are from the STUDIO-F set and test chips are from the STUDIO-M set. ROC area denotes area under the ROC curve. The plot shows ROCs with the top two parameter settings in terms of ROC area for each approach. Detection rates are measured at 2% false positive rate. #levels and #branch are the number of levels and branching factor parameters used to construct the pyramid representation in VGPMK.

Distance measure / kernel	Training time (12k samples)	Testing time (per sample)	# Support vectors
Rigid match	555s	0.0454s	4521
Aligned match	7,020s	0.5532s	5900
VGPMK	22,460s	0.6179s	3140

Table 2. The training time for VGPMK includes construction of the pyramid representation using K-means clustering. LIBSVM was used to train the SVMs for all algorithms. The HOG feature extraction time is not included for all three approaches.

7.3. Hand detection results in cluttered ASL video

To demonstrate hand detection performance using the aligned distance function for the ASL lexicon retrieval application, we use the LAB-F video set of an inexpert signer wearing low skin tone contrast clothing collected in a computer lab. We sample $4,000$ hand chips from STUDIO-F and STUDIO-M sets as foreground examples. We sample $8,000$ background chips from the STUDIO sequences, from a rest pose sequence of our test signer and a lab background sequence as illustrated in Figure 3. We follow the steps as in the previous experiment to extract HOG features and train the SVM based hand detector using a 13 neighborhood aligned distance function. We choose $\gamma = 0.042$ in training the SVM based on results from the previous experiment.

We follow the steps described in Section 5 to detect hand locations in test video sequences. We detect a fixed number of hand candidates in each frame; three candidates are chosen for one handed signs and five candidates are chosen for two handed signs. With ≈ 200 image scan ROIs after skin mask based pruning (Figure 3), 3 and 5 detected hand candidates correspond to false positive rates of $2/200$ and $3/200$ for one and two handed signs respectively. Results of hand detection on example video frames are shown in Figure 6. The total detection time is $\approx$ 100s per frame. Approaches to make the detector more efficient are discussed in the next section.

8. Conclusions and future work

A distance measure is proposed to compute a non-rigid alignment between pairs of hand chips to accommodate hand shape variations for each signer and among different signers. The distance measure is incorporated into a SVM based foreground/background classifier for hand detection. The proposed approach shows better hand detection rates than rigid matching and VGPMK on ASL video of gestures signed by experts. The proposed approach has fewer and easier to tune parameters while being less computationally expensive than VGPMK. Robustness of the proposed approach is demonstrated on video of ASL gestures signed by an inexpert signer in an unconstrained setting with cluttered background.

Techniques to further improve performance of hand detection and part of our future work include,

- ASL constraints: The range of hand shapes within a sign are constrained by ASL production rules. For instance, not every hand shape co-occurs with every other hand shape, and many two handed signs either have symmetric hand shapes or a limited set of hand shapes for the non-dominant hand.
- Clutter model: The signer's face is the most significant contribution to background variation in hand chips, Smith, et al. [19] propose a relevant approach to model facial clutter.
- Forearm detector: A forearm detector can be used to further prune the ROI set for input to the hand detector.
- Regularization term in alignment: A simple mesh model can be used to constrain the non-rigid alignment and smooth the deformation field.

References

[1] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. Boostmap: An embedding method for efficient nearest neighbor retrieval. *IEEE T-PAMI*, 30(1):89–104, 2008. 2

[2] V. Athitsos and S. Sclaroff. Estimating 3D hand pose from a cluttered image. In *CVPR*, 2003. 2

[3] S. Boughorbel, J.-P. Tarel, and N. Boujemaa. The intermediate matching kernel for image local features. In *IJCNN*, 2005. 4

[4] P. Buehler, M. Everingham, D. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language tv broadcasts. In *BMVC*, 2008. 2

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. In *http://www.csie.ntu.edu.tw/~cjlin/libsvm*, 2001. 6

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 2, 3

[7] M. de La Gorce, N. Paragios, and D. J. Fleet. Model-based hand tracking with texture, shading and self-occlusions. In *CVPR*, 2008. 2

[8] K. G. Derpanis, R. P. Wildes, and J. K. Tsotsos. Definition and recovery of kinematic features for recognition of american sign language movements. *Pattern Recognition*, 26:1650–1662, 2008. 2

[9] A. Farhadi, D. Forsyth, and R. White. Transfer learning in sign language. In *CVPR*, 2007. 2

[10] K. Fujimura and L. Xu. Sign recognition using constrained optimization. In *ACCV*, 2007. 2

[11] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005. 2, 3, 4

[12] K. Grauman and T. Darrell. Approximate correspondences in high dimensions. In *NIPS*, 2006. 3, 5

[13] Y. Hamada, N. Shimada, and Y. Shirai. Hand shape estimation under complex backgrounds for sign language recognition. In *AFGR*, 2004. 2

[14] J. J. Lee. LIBPMK: A pyramid match toolkit. Technical Report MIT-CSAIL-TR-2008-17, 2008. 6

[15] H. Ling and S. Soatto. Proximity distribution kernels for geometric context in category recognition. In *ICCV*, 2007. 2, 3, 4

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1, 3

[17] F. Odone, A. Barla, and A. Verri. Building kernels from binary strings for image matching. *IEEE T-IP*, 14(2):169180, 2005. 4

[18] E.-J. Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *AFGR*, 2004. 2

[19] P. Smith, N. da Vitoria Lobo, and M. Shah. Resolving hand over face occlusion. *IVC*, 25:1432–1448, 2007. 3, 7

[20] B. Stenger, P. R. S. Mendonca, and R. Cipolla. Model-based 3D tracking of an articulated hand. In *CVPR*, 2001. 2

[21] R. Tennant and G. Brown. *The American Sign Language Handshape Dictionary*. Gallaudet University Press, 2004. 1

[22] C. Valli, editor. *The Gallaudet Dictionary of American Sign Language*. Gallaudet University Press, 2005. 5

[23] P. Viola and M. Jones. Fast multi-view face detection. In *CVPR*, 2003. 5

[24] J. Wang, V. Athitsos, S. Sclaroff, and M. Betke. Detecting objects of variable shape structure with hidden state shape models. *IEEE T-PAMI*, 30(3):477–492, 2008. 2

[25] Q. Yuan, A. Thangali, V. Ablavsky, and S. Sclaroff. Multiplicative kernels: Object detection, segmentation and pose estimation. In *CVPR*, 2008. 2

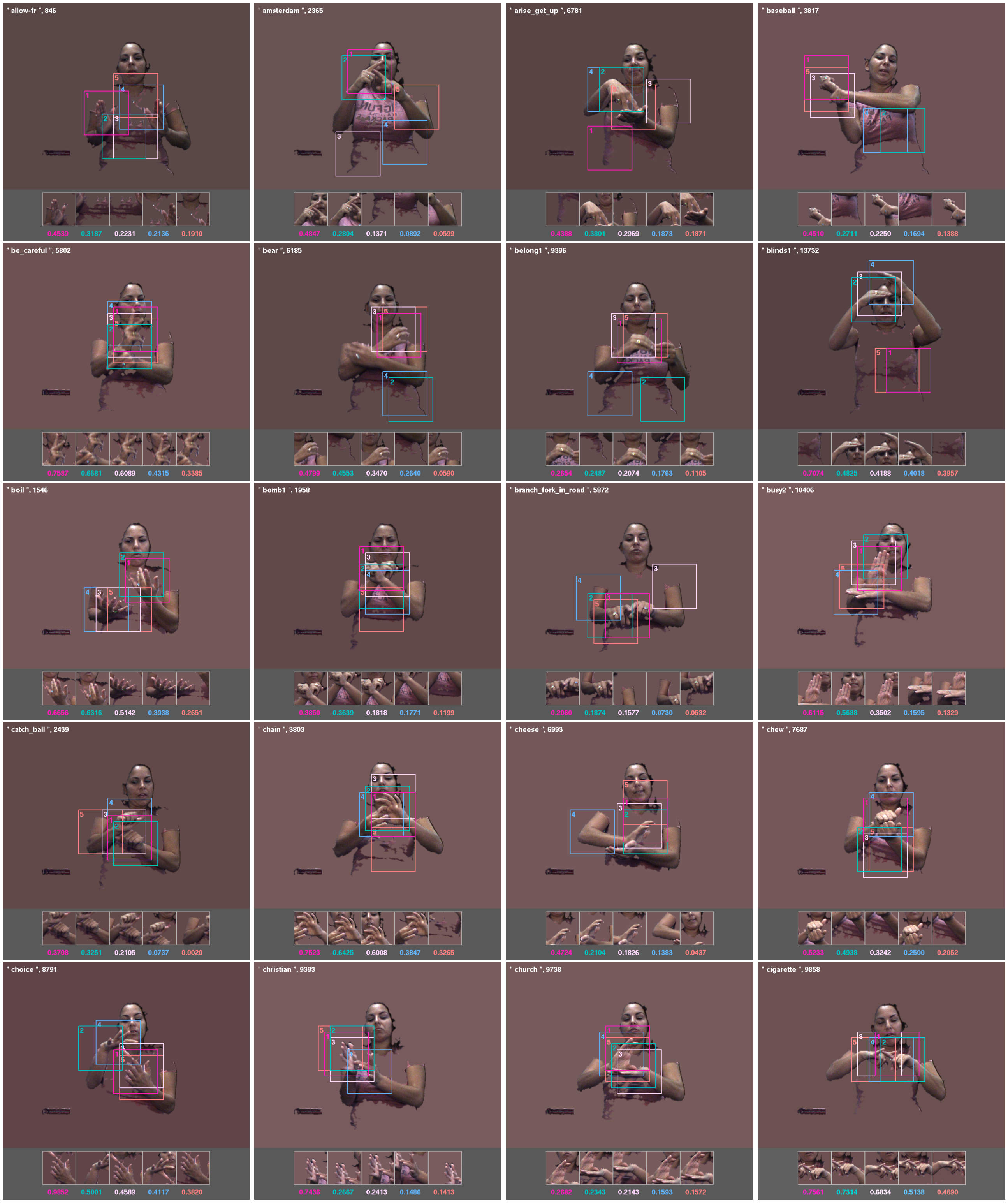

Figure 6. We show hand detection results using the 13 neighborhood aligned distance based detector on gestures with interacting hands from the LAB-F set. The training and test setup for this experiment is described in Section 7.3. The detection scores (i.e., SVM outputs) for top five hand ROIs are displayed for each frame sorted in decreasing order.