Correlation Filters for Object Alignment

Vishnu Naresh Boddeti
Carnegie Mellon University
naresh@cmu.edu

Takeo Kanade
Carnegie Mellon University
tk@cs.cmu.edu

B.V.K. Vijaya Kumar
Carnegie Mellon University
kumar@ece.cmu.edu

Abstract

Alignment of 3D objects from 2D images is one of the most important and well studied problems in computer vision. A typical object alignment system consists of a landmark appearance model which is used to obtain an initial shape and a shape model which refines this initial shape by correcting the initialization errors. Since errors in landmark initialization from the appearance model propagate through the shape model, it is critical to have a robust landmark appearance model. While there has been much progress in designing sophisticated and robust shape models, there has been relatively less progress in designing robust landmark detection models. In this paper we present an efficient and robust landmark detection model which is designed specifically to minimize localization errors thereby leading to state-of-the-art object alignment performance. We demonstrate the efficacy and speed of the proposed approach on the challenging task of multi-view car alignment.

1. Introduction

Fitting a shape or a template to a given image is one of the most important and well studied problems in computer vision where the object shape is typically defined by a set of landmarks. The ability to accurately align shape models of deformable objects is critical for a variety of applications like object detection and recognition, object tracking, 3D scene modeling etc. A typical approach for shape fitting involves a local landmark appearance model which generates a likelihood map for that landmark and a deformable shape model which fits a shape to the landmark likelihood maps. There has been much progress in designing parametrized deformable shape models over the past two decades. In their seminal paper Cootes et al. introduced one of the most successful alignment models, Active Shape Models [5], where the object shape, as represented by a set of landmark points, is modeled by a Gaussian point distribution. As an extension Zhou et.al., [23] proposed a Bayesian Tangent Shape Model (BTSM) where both the object pose and the shape deformation are estimated iteratively using the EM algorithm. In [13] Li et. al. introduced a robust version of BTSM to handle outliers and gross landmark detection errors. Other notable shape matching models include Active Appearance Models [3], Pictorial Structures [10] and Constrained Local Models [6][19].

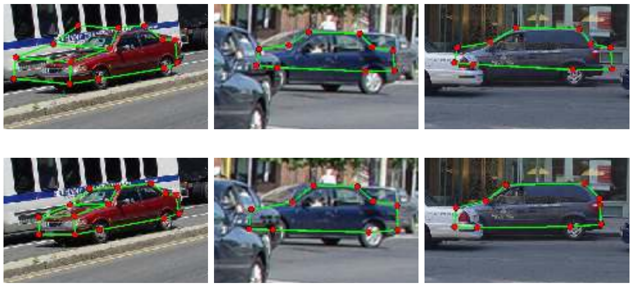


Figure 1. A comparison between two different appearance models used with the same shape model for the task of car alignment. Top Row: car alignment with a random forest based landmark appearance model. Bottom Row: car alignment with the proposed landmark detector.

The goal of the appearance model is to provide an initial shape for the alignment algorithm. Due to background clutter and substantial variations in color and pose, capturing the local appearance can be quite challenging. Discriminative feature representations in conjunction with discriminative classifiers can provide robustness against these challenges. Many different feature representations have been used in the literature for the purpose of landmark detection. Notable examples include Gabor wavelets [12], Haar wavelets [25][4] and Histogram of Oriented Gradients (HOG) [7][24][13] features. The choice of the feature representation used to represent the landmarks critically affects the performance of the landmark detector. We use HOG features to represent the local appearance of the landmarks since they have been known to perform well on a variety of detection tasks [18][9]. Additionally many discriminative appearance models like Support Vector Machines (SVMs) [24][21], RankBoost [22], FisherBoost [20], Random Forests (RFs) [13][4] have been proposed in the literature to detect landmarks in images. However, most of these methods are generic discriminative models i.e., none of them are specifically designed for the task of landmark localization. Correlation filters (CFs) are another class of template based linear classifiers which are designed to minimize localization errors in addition to discriminating between the target object and the background and as such are well suited to the task of landmark detection. While many different correlation filter designs exist [14][2] they have been traditionally designed to be used with scalar feature representations only. In this paper, we present a vector feature based extension to the scalar feature based unconstrained correlation filters [15][1]. The proposed vector correlation filter (VCF) with HOG features can accommodate significant within-class variations while being discriminative against background clutter. We demonstrate that VCFs are fast and accurate landmark detectors and can provide robust object alignment when used in conjunction with point based shape models (see Fig. 1 for a comparison between a VCF and RF based landmark detector for the same shape model). We also show that VCFs outperform other popular landmark detectors based on SVMs and RFs leading to state of the art object alignment accuracy.

2. Proposed Approach

Traditionally the focus of most research on object alignment models has been on improving the shape models. Many shape models have been proposed to account for and be robust to ever larger errors made by the appearance models. However, large gains in object alignment performance can also be had from designing more robust appearance models which also results in the shape models having to contend with lower noise levels, thereby improving their performance. We now describe the proposed landmark appearance model and the shape model that we use for the task of object alignment.

2.1. Appearance Model

The primary task of the appearance model is to serve as an initialization for the landmark search and as such requires very good localization performance. Although RFs and SVMs are widely used to design discriminative appearance models, they are not explicitly designed for localization and hence are unable to provide high localization accuracy. Correlation filters (CFs) are another class of classifiers which are generally designed for high localization performance and are hence better suited to the task of landmark detection. We briefly describe CFs for scalar features before we introduce their vector formulation.

2.1.1 Correlation Filters

A CF is a spatial-frequency array (equivalently, a template in the image domain) that is specifically designed from a set of training patterns that are representative of a particular pattern class. This template is compared to a query image by obtaining the cross-correlation as a function of the relative shift between the template and the query. For computational efficiency this is computed in the spatial frequency domain (u,v), i.e.,

$$C(u,v) = I(u,v)F^{*}(u,v) \tag{1}$$

where $I(u,v)$ is the 2D Fourier transform (FT) of the query pattern and $F(u,v)$ is the CF (i.e., 2D FT of the template) and $C(u,v)$ is the 2D FT of the correlation output $c(x,y)$ with superscript $*$ denoting the complex conjugate. Since the images and their FTs are discrete-indexed, FT here refers to the discrete Fourier transform (DFT) which is implemented via the Fast Fourier Transform algorithm (FFT). The CFs are usually designed to give a sharp peak at the center of the correlation output plane $c(x,y)$ for a centered authentic query pattern and no such peak for an impostor.

The main idea behind correlation filters is to control the shape of the cross-correlation output between the image and the filter by minimizing the average Mean Square Error (MSE) between the cross-correlation output and the ideal desired correlation output for an authentic (or impostor) input image. By explicitly controlling the shape of the entire correlation output, unlike traditional classifiers which only control the output value at the target location, CFs achieve more accurate target localization. For N training images the CF design problem is posed as an optimization problem (for notational ease expressions are given for 1-D signals),

$$\min_{\mathbf{f}} \frac{1}{N}\sum_{i=1}^{N} \left\|\mathbf{x_i}\otimes\mathbf{f}-\mathbf{g_i}\right\|_2^2 + \lambda\left\|\mathbf{f}\right\|_2^2 \tag{2}$$

where $\otimes$ denotes the convolution operation, $\mathbf{x_i}$ denotes the $i-$th image, $\mathbf{f}$ is the CF template and $\mathbf{g_i}$ is the desired correlation output for the $i-$th image and λ is the regularization parameter. This optimization problem can be solved efficiently in the frequency domain where the objective function has the following closed form expression,

$$\min_{\hat{\mathbf{f}}} \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{f}}^{\dagger}\hat{\mathbf{X}}_{\mathbf{i}}^{\dagger}\hat{\mathbf{X}}_{\mathbf{i}}\hat{\mathbf{f}} - \frac{2}{N}\sum_{i=1}^{N}\hat{\mathbf{f}}_{\mathbf{i}}^{\dagger}\hat{\mathbf{X}}_{\mathbf{i}}^{\dagger}\hat{\mathbf{g}}_{\mathbf{i}} + \lambda\hat{\mathbf{f}}^{\dagger}\hat{\mathbf{f}} \tag{3}$$

where $\hat{\mathbf{x}}$ denotes the Fourier transform of $\mathbf{x}$ and $\hat{\mathbf{X}}$ denotes the diagonal matrix whose diagonal entries are the elements of $\hat{\mathbf{x}}$ and $\dagger$ denotes conjugate transpose. Solving the above optimization problem results in the following closed form expression for the CF,

$$\hat{\mathbf{f}} = \left[\lambda\mathbf{I} + \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{X}}_{\mathbf{i}}^{\dagger}\hat{\mathbf{X}}_{\mathbf{i}}\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{X}}_{\mathbf{i}}^{\dagger}\hat{\mathbf{g}}_{\mathbf{i}}\right] \tag{4}$$

where $\mathbf{I}$ is the identity matrix of appropriate dimensions.

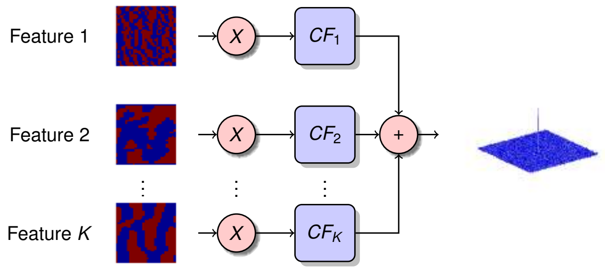


Figure 2. Vector Correlation Filter: The outputs of each feature channel are aggregated to compute the final correlation output which would have a sharp peak at the target location.

2.1.2 Vector Correlation Filter

Traditional correlation filters have been often designed using scalar features (most commonly pixel values) and cannot be directly used with vector features like HOG. We propose an extension of the unconstrained correlation filters described in Eq. 2 to vector valued features (like HOG with $K = 36$ feature channels). The VCF, consists of one correlation filter per feature channel which are all jointly optimized to minimize the localization loss defined as the l_2-norm of the difference between the correlation output and the desired ideal correlation output. Since each feature (corresponding to each branch, see Fig. 2 for a pictorial description of VCFs) leads to a peak (at least for the correct object) at the same location the final output can be obtained by coherently adding all the branch outputs. Therefore the filter design problem can be formulated as,

$$\min_{\mathbf{f^1},\mathbf{f^2},\ldots,\mathbf{f^K}} \frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{k=1}^{K}\mathbf{x_i^k}\otimes\mathbf{f^k}-\mathbf{g_i}\right\|_2^2+\lambda\sum_{k=1}^{K}\left\|\mathbf{f^k}\right\|_2^2 \quad (5)$$

We now pose this minimization problem equivalently in the frequency domain (using Parseval's Theorem [17]) to derive a closed form expression which in turn lends itself to an efficient solution.

$$\min_{\mathbf{\hat{f}^1},\mathbf{\hat{f}^2},\ldots,\mathbf{\hat{f}^K}} \frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{k=1}^{K}\mathbf{\hat{X}_i^k}\mathbf{\hat{f}^k}-\mathbf{\hat{g}_i}\right\|_2^2+\lambda\sum_{k=1}^{K}\left\|\mathbf{\hat{f}^k}\right\|_2^2 \quad (6)$$

The optimization problem in Eq. 6 can be reduced to the following unconstrained quadratic minimization problem,

$$\min_{\mathbf{\hat{f}}} \quad \mathbf{\hat{f}}^\dagger\mathbf{\hat{S}}\mathbf{\hat{f}}-2\mathbf{\hat{f}}^\dagger\mathbf{\hat{r}} \quad (7)$$

where $\mathbf{\hat{S}} = \mathbf{\hat{D}} + \lambda\mathbf{I}$, with $\mathbf{I}$ being an identity matrix of appropriate dimensions, and

$$\mathbf{\hat{D}} = \begin{bmatrix} \frac{1}{N}\sum_i \mathbf{\hat{X}_i^{1}}^\dagger\mathbf{\hat{X}_i^1} & \cdots & \frac{1}{N}\sum_i \mathbf{\hat{X}_i^{1}}^\dagger\mathbf{\hat{X}_i^k} \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_i \mathbf{\hat{X}_i^{k}}^\dagger\mathbf{\hat{X}_i^1} & \cdots & \frac{1}{N}\sum_i \mathbf{\hat{X}_i^{k}}^\dagger\mathbf{\hat{X}_i^k} \end{bmatrix} \quad (8)$$

$$\mathbf{\hat{r}} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}\mathbf{\hat{X}_i^{1}}^\dagger\mathbf{\hat{g}_i} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N}\mathbf{\hat{X}_i^{k}}^\dagger\mathbf{\hat{g}_i} \end{bmatrix} \quad \mathbf{\hat{f}} = \begin{bmatrix} \mathbf{\hat{f}^1} \\ \vdots \\ \mathbf{\hat{f}^k} \end{bmatrix} \quad (9)$$

where $\mathbf{D}$ is the cross-power spectrum matrix (interaction energy between the feature channels). The parameter, λ, offers a trade-off between the localization loss and the l_2-regularization. The VCF is a powerful detector that uses many degrees of freedom to satisfy the design criteria for detecting similarities between members of the same pattern class. In contrast to traditional CF designs and SVMs, which treat each feature channel as being independent of each other, the VCF design jointly optimizes the performance of multiple channels to produce the desired output plane by taking advantage of the joint properties of the different feature channels via interactions between the multiple feature channels. Solving for $\mathbf{\hat{f}}$ in Eq. 7 results in the following closed form expression for the VCF,

$$\mathbf{\hat{f}} = \left[\lambda\mathbf{I}+\mathbf{\hat{D}}\right]^{-1}\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}\mathbf{\hat{X}_i^{1}}^\dagger\mathbf{\hat{g}_i} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N}\mathbf{\hat{X}_i^{k}}^\dagger\mathbf{\hat{g}_i} \end{bmatrix} \quad (10)$$

Note that unlike in Eq. 4 where the matrix to be inverted is a diagonal matrix, the matrix to be inverted in Eq. 10 is a non-diagonal matrix with a block matrix structure where each block is a diagonal matrix. Naively inverting the matrix in Eq. 10 is computationally intensive, however by taking advantage of the unique nature of the matrix, it can be inverted quite efficiently by a block-wise matrix inversion. During test time the filter is applied by convolving each feature channel filter with its corresponding feature channel and finally summing up all the feature channel outputs. For computational efficiency the convolutions are performed in the frequency domain.

2.2. Shape Model

Most approaches model the statistical distribution of the shape parameters so that the observed shape from local landmark detectors can be regularized by a prior model during image fitting. The proposed appearance model can be used in conjunction with any of these shape models. For our purposes we use the robust shape model introduced in [13] by Li et.al. due to its ability to handle gross landmark detection errors caused either by partial occlusions or clutter in the background. This model works on the premise that while the object shape is described by multiple landmark points, the actual shape lies in a low-dimensional subspace. Therefore, a small minimal subset of uncorrupted landmarks are sufficient to estimate and hallucinate the full shape (via a Bayesian Partial Shape Inference (BPSI) algorithm, see [13] for details). They adopt a hypothesis and test approach by performing a combinatorial search over the space of possibly occluded landmarks. This explicit search results in a very robust shape alignment model which performs well under all conditions. The original algorithm however ignores the landmark confidence from the appearance model and generates many random partial shapes via Random Sample Consensus (RANSAC) [11] resulting in the evaluation of a very large number of hypotheses for a given probability of sampling a "good" subset. However, by generating the subsets to include landmarks with high confidence, fewer hypotheses can be evaluated to find a "good" subset with high probability. Therefore we propose a modification where we generate all $\binom{n}{k}$ subsets of size k out of the top n confident landmarks instead of choosing random subsets of size k. Through the rest of this paper we refer to the original model as "RANSAC BPSI" and the modified model as "Greedy BPSI". The final object shape is determined by evaluating these hypotheses and choosing the hypothesis with the minimum error between hallucinated shape and the observed shape. The resulting estimate can be further refined by including more inliers and re-estimating the object shape. Mathematically the full BPSI model is defined as,

$$\begin{cases} S = \Phi b + \mu + \epsilon \\ Y_p = M_p(sRS + t + \eta) \\ Y_h = M_h(sRS + t) \end{cases}$$

where μ is the mean shape, Φ is an eigen-vector matrix, ϵ and η account for noise, S is the latent canonical object shape, $\Theta = \{s, R, t\}$ are the pose parameters, M_p is the mask to extract the partial shape, M_h is the mask complementary to M_p, Y_p is the partial shape and Y_h is the complement of Y_p. Inference is done using an EM-algorithm to estimate the model parameters $\Pi = \{b, \Theta\}$ iteratively. In the E-step the posterior of S is computed give the partial observation Y_p and $\Pi^{(n-1)}$ and in the M-step the model parameters $\Pi^{(n)}$ are optimized to maximize the expectation of the data log-likelihood $\log p(Y_p, S|\Pi^{(n)})$ over the missing data posterior $p(S|Y_p, \Pi^{(n-1)})$.

3. Experiments

To evaluate the efficacy of the proposed approach, we consider the task of multi-view car alignment from a single image. This is a challenging task since most car parts are only weakly discriminative for detection and the appearance of the cars can change dramatically as the viewing angle changes. Further cars in natural street scenes vary widely in shape and are often present in highly cluttered backgrounds, with severe occlusion, self or otherwise, in many instances.

3.1. Database

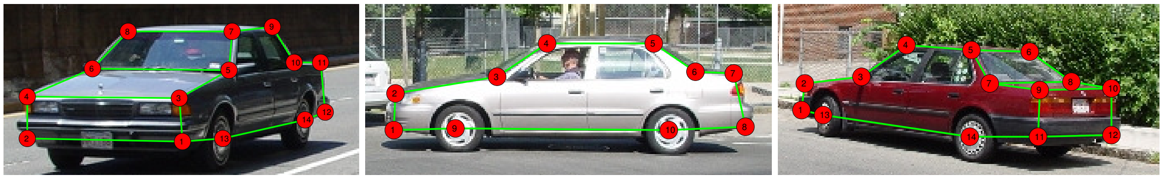
We evaluate the proposed approach on cars from the MIT Street Dataset [16] which contains over 3500 street scene images created for the task of object recognition and scene understanding. This dataset has annotated landmarks (available at `www.cs.cmu.edu/~vboddeti/alignment.html`) for 3,433 cars spanning a wide variety of types, sizes, backgrounds and lighting conditions including partial occlusions. All the shapes are normalized to roughly a size of 250×130 by Generalized Procrustes Analysis [8]. The car shape is represented by 8, 14, 10, 14 and 8 landmarks (see Fig. 3) respectively. The labeled data was further manually classified into five different views: 932 frontal view, 1400 half-frontal view, 803 profile view, 1230 half-back view and 1162 back view images. Due to space constraints we report results only on the half-frontal, profile and half back view. We randomly selected 400 images from each view for training and use the rest of the images for testing. Patches from occluded landmarks are excluded while training the appearance model and for evaluation the occluded landmark is placed at the most likely location in the image.

3.2. Training

For each landmark, we extract a 96×96 image patch as the positive sample and negative samples of the same size are extracted uniformly around each landmark. Each of these local patches are further represented by the Histogram of Oriented Gradients (HOG) descriptor. The HOG descriptors are computed over dense and overlapping grids of spatial blocks, with image gradient features extracted at 9 orientations and a spatial bin size of 4×4. The RF (using the author's implementation from [13]), Linear SVM and VCF are designed using these HOG representations of the patches. To design the VCF, we set the desired correlation output $\mathbf{g}$ for the positive samples to a positively scaled Gaussian at the patch center and to a negatively scaled Gaussian for the negative samples i.e., $g_i(x, y) = t_i \exp\left(-\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2\sigma^2}\right)$ where $t_i = 1$ for a positive patch and $t_i = -0.1$ for a negative patch and (μ_x, μ_y) is the location of the landmark in the patch.

3.3. Evaluation

Quantitatively we evaluate the performance of each car alignment algorithm by computing the root mean square error (RMSE) of the detected landmarks with respect to manually labeled ground truth landmark locations. More specifically we report the landmark-wise average RMSE over four different subsets, 1) average over all images , 2) average over images with no occluded landmarks, 3) average over the unoccluded landmarks in partially occluded images and 4) average over the occluded landmarks in partially occluded images. Fig. 7 shows qualitative alignment results on some challenging images.



(a) View 2 (b) View 3 (c) View 4

Figure 3. Ground truth landmarks with labeled landmark indices for different viewpoints.

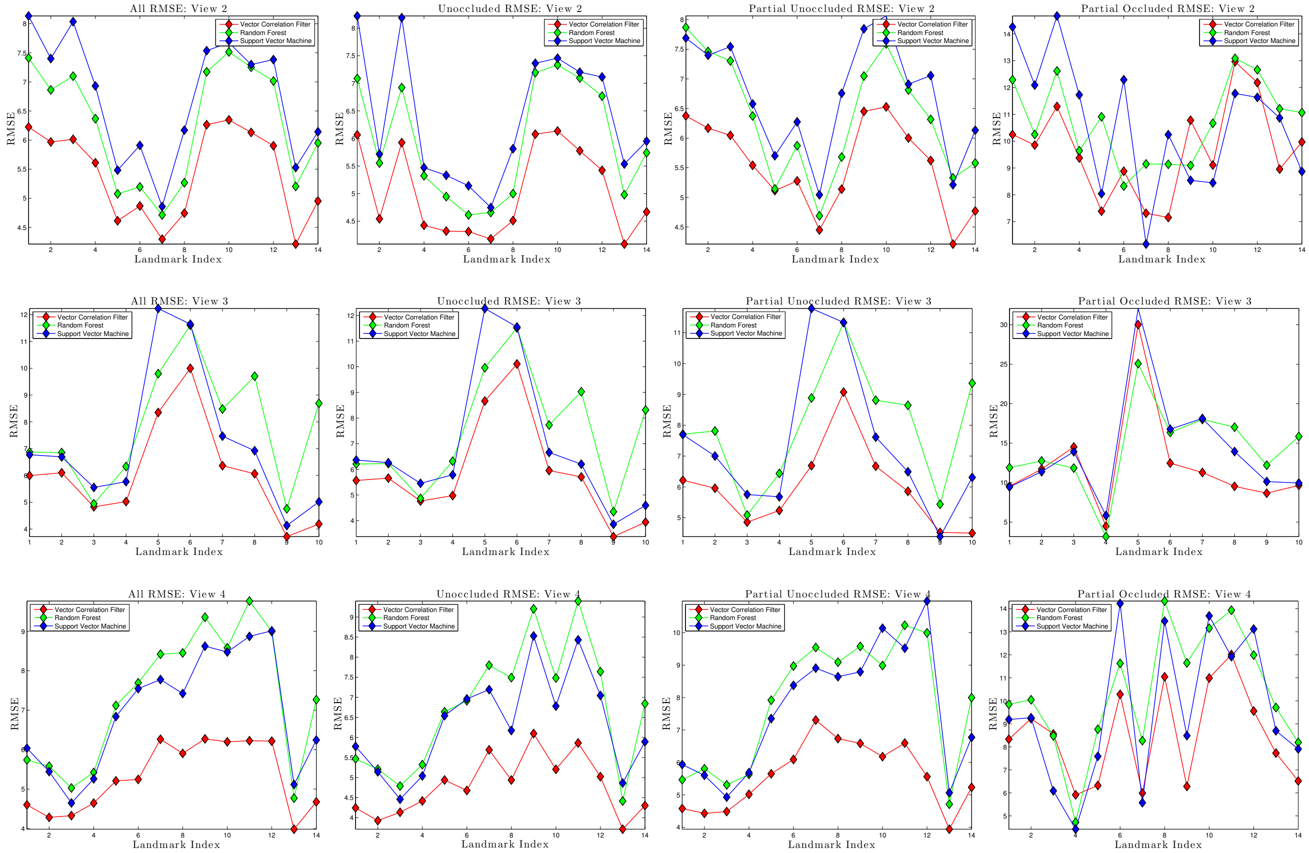


Figure 4. RMSE for each pose averaged over 1) all images, 2) images with no occlusions, 3) unoccluded landmarks of partially occluded images, 4) occluded landmarks of partially occluded images. We compare between three different appearance models, Vector Correlation Filter, Random Forests and Linear SVM, all with the RANSAC BPSI shape model.

3.3.1 Comparing Appearance Models

We compare three different discriminative appearance models for detecting landmarks namely, RFs, Linear SVMs and the proposed landmark detector, VCF. Fig. 4 shows the landmark wise average error for three different poses. Our alignment results with the VCF are a significant improvement over the previous state-of-the-art results [13] (RFs based appearance model that we compare against) on this dataset. We observe that the VCF consistently results in lower RMSE over the other appearance models especially over the unoccluded landmarks where the appearance model directly influences the final result. Since the landmark detection performance over the unoccluded landmarks also indirectly influences the hallucinated landmark locations of the occluded landmarks, we observe a lower RMSE even for occluded landmarks on account of the detector's better performance on the unoccluded landmarks in the partially occluded car images. To further investigate the error distribution we plot the individually sorted errors for each pose in Fig. 5 with the x-axis representing the alignment difficulty for each detector. Notice that for a given error tolerance the VCF aligns more images compared to RFs and SVMs. Comparing the performance of VCF and RF on a per image basis, VCF results in lower RMSE (cumulative

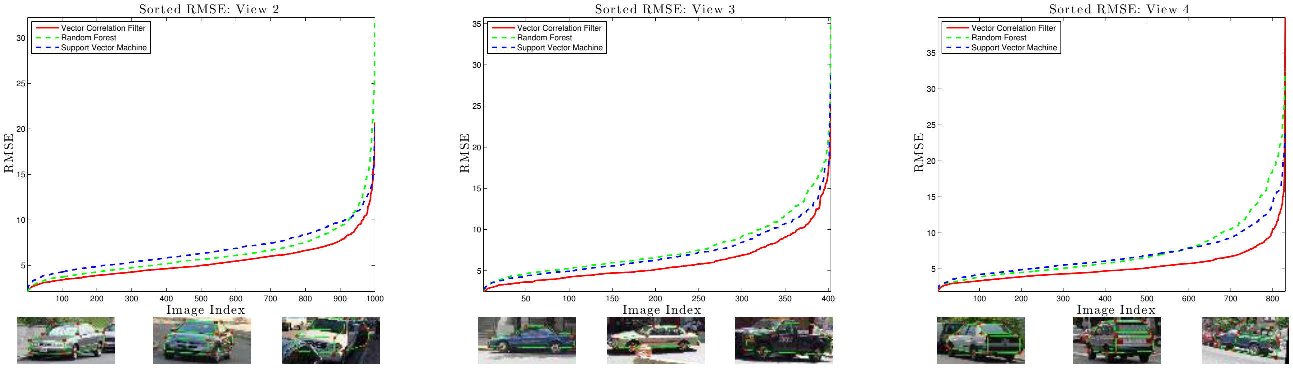


Figure 5. Comparison of the sorted RMSE for each pose for different appearance models along with example alignment results with the VCF appearance model corresponding to small, medium and large landmark RMSE.

RMSE over all the landmarks in the image) for 649 images in comparison to RF and 804 images in comparison to SVM out of 1000 images in view 2.

3.3.2 Comparing Shape Models

Here we compare different shape models for detecting landmarks namely, RANSAC BPSI and its modification Greedy BPSI. We omit a comparison to other shape models like ASM and BTSM due to space constraints and we observe that RANSAC BPSI outperforms both ASM and BTSM which is consistent with the observations made in [13]. Further to evaluate the robustness of RANSAC BPSI and Greedy BPSI to occlusions, we evaluate BPSI assuming that the landmark occlusion masks are known. We refer to this as "Oracle BPSI" through the rest of the paper. Fig. 6 shows the landmark-wise average error for three different poses. We make the following observations, 1) knowing that a landmark is occluded helps as is evident from the RMSE of occluded landmarks on vehicles with partial occlusions (compare "Oracle BPSI" and "RANSAC BPSI"), 2) in addition to treating occluded landmarks as occluded, it is also beneficial to ignore any low confidence landmarks and treat them as occluded landmarks. This is evident from the RMSE of the unoccluded landmarks and the unoccluded landmarks on partially occluded cars (compare "Oracle BPSI" and "RANSAC BPSI" or "Greedy BPSI"), 3) Greedy BPSI outperforms RANSAC BPSI on the occluded landmarks of partially occluded vehicles. This is because Greedy BPSI is more likely to omit the occluded landmarks for estimating the shape due to lower confidence from the appearance model while RANSAC BPSI, which selects the landmarks randomly, is more likely to include occluded landmarks in the hypothesized subset. Moreover, RANSAC BPSI and Greedy BPSI perform equally well on images with unoccluded landmarks. Comparing the performance of RANSAC BPSI and Greedy BPSI on a per image basis for 1000 images in view 2, Greedy BPSI results in lower RMSE (cumulative RMSE over all the landmarks in the image) for 520 images (i.e., lower RMSE on 40 images) in comparison to RANSAC BPSI and 530 images (i.e., lower RMSE on 60 images) in comparison to Oracle BPSI, 4) finally, analysis of the alignment results reveals that the model performs well on sedan like vehicle and most of the errors are on vehicles like jeeps, trucks and vans due to an overwhelming majority of cars in the dataset being sedans. Due to the considerable difference in the shape of these vehicles a mixture model with different shape models for different car types (e.g., pickup trucks, vans, sedans, jeeps etc.) can help further improve alignment performance.

4. Computational Complexity

Table 1. Execution Time (in ms) Per Image on Single Core

Pose	RF	VCF/SVM	RANSAC BPSI	Greedy BPSI
2	4000	200	700	90
3	3000	150	600	70
4	4000	200	700	90

Table 1 shows the timing results for the appearance and shape model individually for a C++ based implementation on a 2.7GHz laptop with 4GB RAM. Assuming that the VCFs are stored in the frequency domain, given a new image, the computational complexity for a MxN image with K channels, L landmarks and C templates per landmark is given by, $T = K{*}T_{FFT}+L{*}C{*}T_{FFT}+\mathcal{O}(K{*}L{*}C{*}MN)$ where $T_{FFT} = \mathcal{O}(MN * log_2 MN)$. By making use of the fact that images are real and that the output of the appearance model is real, one can decrease memory usage and computations by a factor of 2. Further Greedy BPSI evaluates 56 hypothesis ($n = 8$ and $k = 5$) while RANSAC BPSI evaluates about 450 hypothesis (for a desired probability of success $p = 0.99$), resulting in a 8x speedup of the shape model without any loss in alignment accuracy. We note that linear SVMs have the exact computational com-

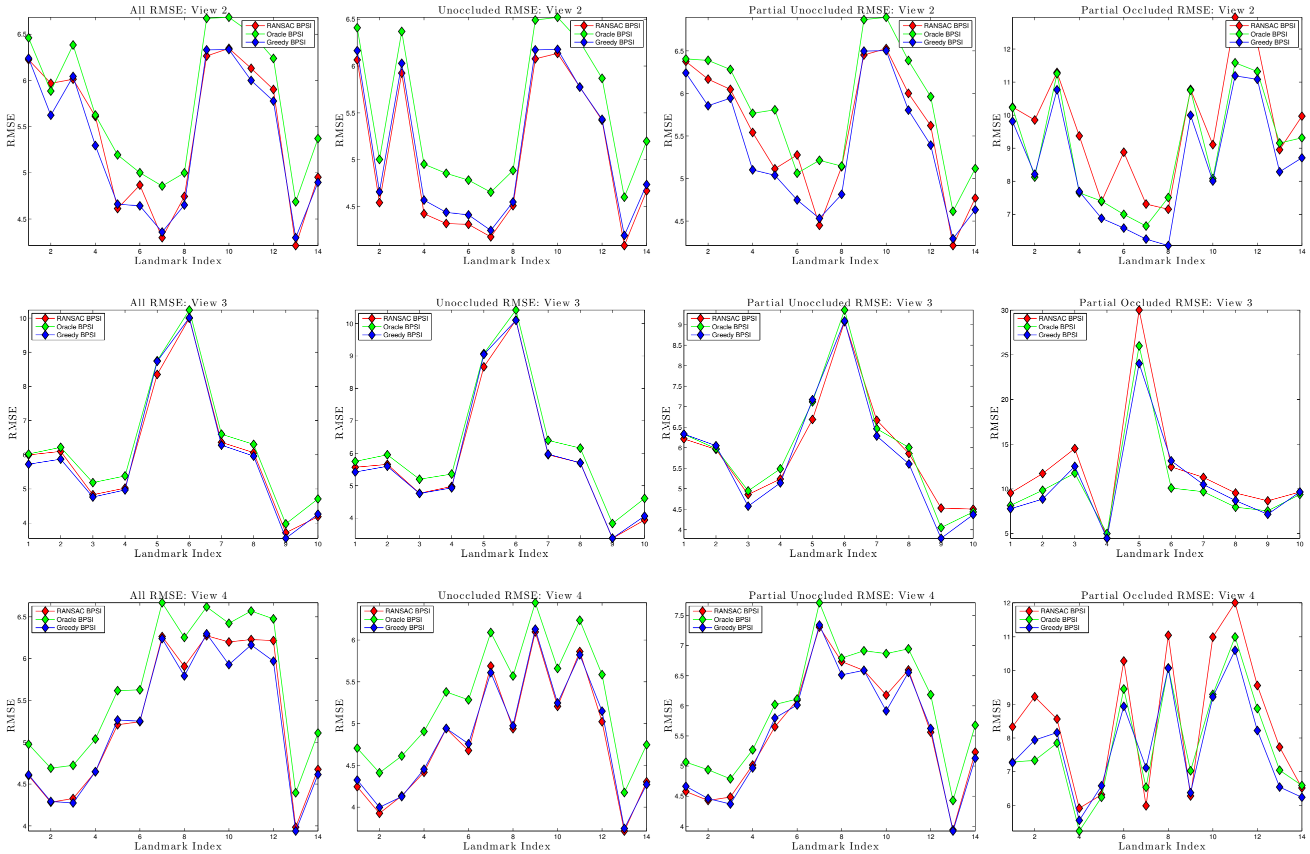


Figure 6. RMSE for each pose averaged over 1) all images, 2) images with no occlusions, 3) unoccluded landmarks of partially occluded images, 4) occluded landmarks of partially occluded images. We compare between three different shape models, RANSAC BPSI, BPSI with occlusion oracle and Greedy BPSI, all with the VCF appearance model.

plexity as VCF, while RFs are slower since every window needs to be scanned explicitly.

5. Conclusion

High accuracy object shape alignment, requires a high accuracy landmark detector as well as a robust shape model. While much work has been done on designing robust shape models, there has been lesser progress on designing robust landmark detection models. In this paper we proposed a robust and fast landmark detector which is specifically designed to minimize localization errors. On the challenging task of multi-view car alignment we observed a significant improvement in the alignment accuracy.

References

[1] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2544–2550, 2010. 2

[2] D. S. Bolme, B. A. Draper, and J. R. Beveridge. Average of synthetic exact filters. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2105–2112, 2009. 2

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. 1

[4] T. Cootes, M. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. *European Conf. on Computer Vision 2012*, pages 278–291, 2012. 1

[5] T. Cootes, C. Taylor, D. Cooper, J. Graham, et al. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 1

[6] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008. 1

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 886–893, 2005. 1

[8] I. Dryden and K. Mardia. *Statistical Shape Analysis*. John Wiley Sons, 1998. 4

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 1

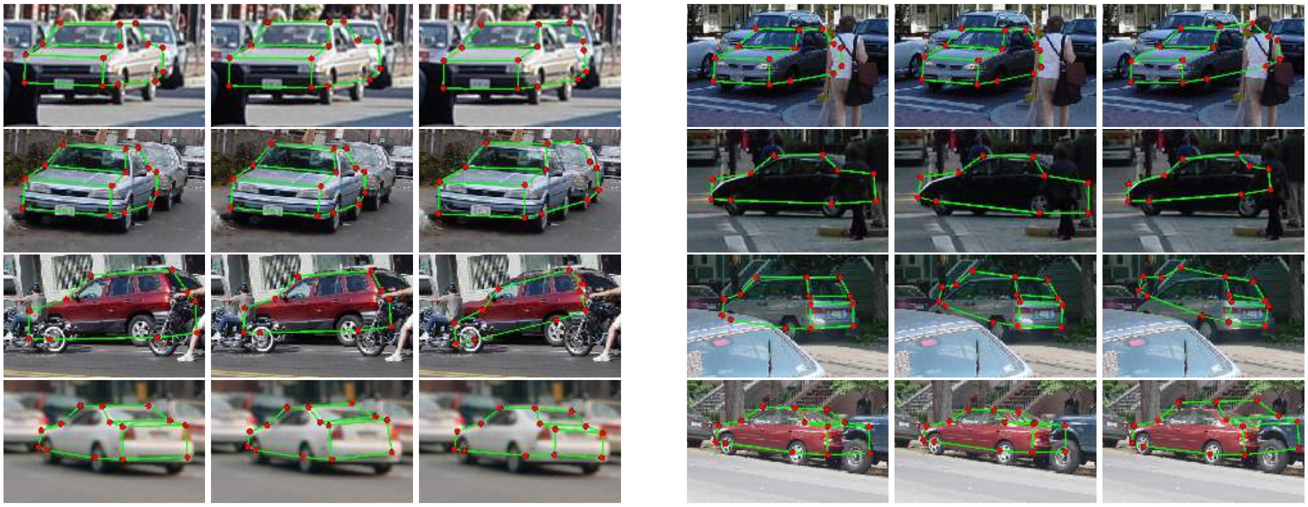


Figure 7. Car Alignment: The three columns correspond to 1) VCF + Greedy BPSI, 2) VCF + RANSAC BPSI, 3) RF + RANSAC BPSI

[10] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int'l Journal of Computer Vision*, 61(1):55–79, 2005. 1

[11] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4

[12] F. Jiao, S. Li, H. Shum, and D. Schuurmans. Face alignment using statistical models and wavelet features. In *IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages I–321, 2003. 1

[13] Y. Li, L. Gu, and T. Kanade. A robust shape model for multiview car alignment. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2466–2473, 2009. 1, 3, 4, 5, 6

[14] A. Mahalanobis, B. V. K. Vijaya Kumar, and D. Casasent. Minimum average correlation energy filters. *Applied Optics*, 26(5):3633–3640, 1987. 2

[15] A. Mahalanobis, B. V. K. Vijaya Kumar, S. Song, S. Sims, and J. Epperson. Unconstrained correlation filters. *Applied Optics*, 33(17):3751–3759, 1994. 2

[16] MIT-StreetScene. http://cbcl.mit.edu/software-datasets/streetscenes/. http://cbcl.mit.edu/software-datasets/streetscenes/. 4

[17] A. V. Oppenheim, A. S. Willsky, and S. Hamid. *Signals and Systems*. Prentice Hall, 1997. 3

[18] J. Saragih. Principal regression analysis. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2881–2888, 2011. 1

[19] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *Int'l Journal of Computer Vision*, 91(2):200–215, 2011. 1

[20] J. Tu, Z. Zhang, Z. Zeng, and T. Huang. Face localization via hierarchical condensation with fisher boosting feature selection. In *IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, pages II–719, 2004. 1

[21] Y. Wang, S. Lucey, and J. Cohn. Enforcing convexity for improved alignment with constrained local models. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–8, 2008. 1

[22] S. Yan, M. Li, H. Zhang, and Q. Cheng. Ranking prior likelihood distributions for bayesian shape localization framework. In *IEEE Int'l Conf. on Computer Vision*, pages 51–58, 2003. 1

[23] Y. Zhou, L. Gu, and H. Zhang. Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference. In *IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages I–109. IEEE, 2003. 1

[24] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2879–2886, 2012. 1

[25] F. Zuo et al. Fast facial feature extraction using a deformable shape model with haar-wavelet based local texture attributes. In *Int'l Conf. Image Processing*, volume 3, pages 1425–1428, 2004. 1