# Multi-Camera Head Pose Estimation Using an Ensemble of Exemplars

Scott Spurlock, Peter Malmgren, Hui Wu, and Richard Souvenir
Department of Computer Science
University of North Carolina at Charlotte
9201 University City Blvd., Charlotte, NC 28223
{sspurloc, ptmalmyr, hwu13, souvenir}@uncc.edu

## ABSTRACT

We present a method for head pose estimation for moving targets in multi-camera environments. Our approach utilizes an ensemble of exemplar classifiers for joint head detection and pose estimation and provides finer-grained predictions than previous approaches. We incorporate dynamic camera selection, which allows a variable number of cameras to be selected at each time step and provides a tunable trade-off between accuracy and speed. On a benchmark dataset for multi-camera head pose estimation, our method predicts head pan angle with a mean absolute error of $\sim 8^\circ$ for different moving targets.

## CCS Concepts

•**Computing methodologies** → **Computer vision tasks;** *Tracking; Ensemble methods;*

## Keywords

exemplar-based learning; distributed cameras; head pose

## 1. INTRODUCTION

Head pose provides cues to a subject's attention and focus, which can be important for applications in surveillance, marketing, and HCI. Multi-camera networks are well-suited to support these applications; however, in the most common deployments, cameras observe a wide field-of-view, and a person only occupies a small area of image, with heads sometimes as small as 20 pixels. In addition, the motion of people in the scene introduces challenges due to changes in scale and perspective.

In this paper, we propose a method for head pose estimation in multi-camera networks that is based on an ensemble of exemplars, which can be used to build a strong predictor using relatively simple features. We introduce a dynamic camera selection scheme, which allows the system to use the prediction from fewer cameras in "easy" cases (e.g., large faces, visible facial features) and more cameras in cases of ambiguity. Our main contributions are (1) adapting exemplar classification to the problem of head pose estimation, (2) providing fine-grained predictions of head pose angles, and (3) dynamic camera selection to balance accuracy and computational efficiency.



*ICDSC '15, September 08 - 11, 2015, Seville, Spain*

 ISBN 978-1-4503-3681-9/15/09…$15.00

DOI: http://dx.doi.org/10.1145/2789116.2789123

**Figure 1: Our method uses an ensemble of exemplars to provide fine-grained head pose estimates in multi-camera networks.**

## 2. RELATED WORK

Head pose estimation is often used as a proxy for gaze estimation [10]. In the cases where facial features are readily identifiable from images, gaze direction can be estimated directly via eye detection and pupil tracking [6]. At medium-scale resolutions, some approaches rely on locating salient features such as the eyes, ears, and nose [18]. Several recent efforts have sought to estimate head pose relative to a single camera from low-resolution images. One approach introduced a descriptor based on Kullback-Leibler distance applied to facial appearance [11]. Tosato et al. describe a new feature descriptor (ARCO) targeted at vision tasks in low-resolution images [14].

For the multi-camera setting, approaches tend to fall into one of several categories. Some work has investigated synthesizing 3D head shapes, e.g., ellipsoids [1] or spheres [15]. These methods tend to be computationally expensive and require many cameras. Other methods concatenate images from network cameras to learn a single discriminative function [7]. The most common approach applies existing monocular head pose techniques separately to individual views, computing relative pose (or a probability distribution of relative pose) for each camera and combining to compute the

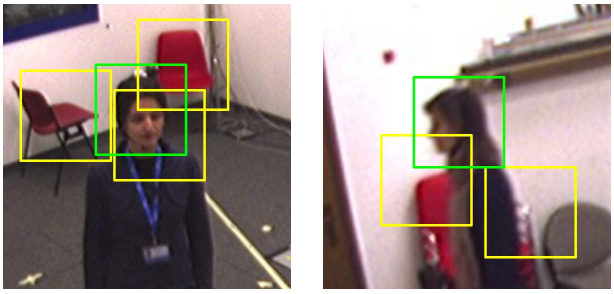


**Figure 2: For joint localization and pose estimation, multiple patches are evaluated at each camera. The highest-scoring detection is shown in green.**

absolute pose estimate (e.g., [9]). Unlike these methods, our is applicable to the case of moving targets.

Some work has addressed head pose estimation of moving people in multi-camera settings. Yan et al. proposed a scheme to learn head appearance as a function of position within an environment [16, 17]. Other methods have incorporated transfer learning to leverage information from datasets with stationary people [12]. These approaches provide coarse predictions of head pose as one of a small number of pre-defined directions. Our method is designed for continuous pose estimation for moving people in low-resolution images. Further, it has low computational cost, using neither complex features nor expensive model fitting.

# 3. METHOD

The focus of this paper is to estimate the head pan (azimuth) and tilt (elevation) angles with respect to a global coordinate system in calibrated, multi-camera networks. This is normally one step in a pipeline that includes detection and tracking, so we assume that the target has been localized (e.g., bounding box in each camera). This leaves the problems of head localization and pose estimation. Previous approaches consider these two issues separately and often employ computationally expensive methods for head localization [12]. In this section, we describe our computationally-efficient, joint approach to head localization and pose estimation from a single camera, and also the aggregation scheme for multi-camera networks.

## 3.1 Single-Camera Head Pose Estimation

Previous work has shown that most features used for head pose estimation are sensitive to localization, especially in the case where the targets move freely [17]. Given a tracked target in a multi-camera network, a rough localization of the head can be obtained using simple rules (e.g., top third of of the target's bounding box). We use a sliding window approach, as shown in Figure 2, for evaluating multiple locations and a prediction scheme that provides a confidence level associated with the prediction.

### 3.1.1 Exemplar SVM

For joint localization and estimation, we use an ensemble of exemplar SVMs (ESVM), which has been previously applied to object detection [8] and place recognition [4]. Figure 3 provides a visual overview of ESVM, where local detectors are trained using a single (positive) exemplar. Figure 4 shows positive and negative image patches for the problem

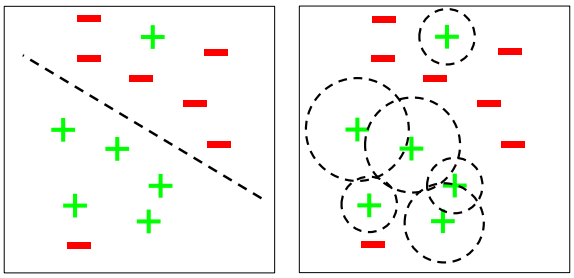


**Figure 3: (Left) Most learning methods fit a global model to the training data. (Right) Exemplar SVMs learn local models centered on individual exemplars. The ensemble of the local learners represents a complex prediction model.**

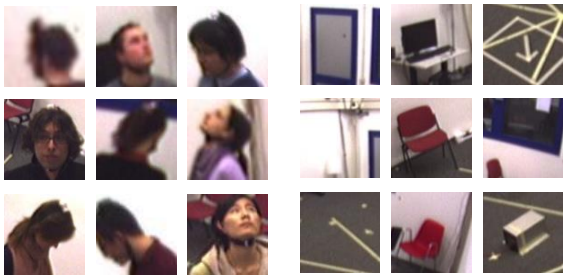


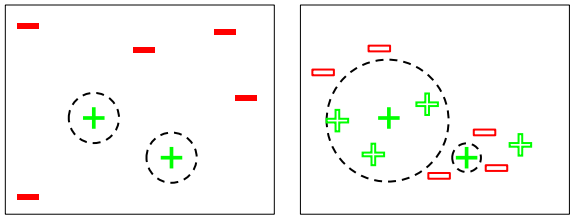
**Figure 4: Exemplar SVM models are trained using a single positive training example (left) and many negative examples (right).**

of head pose estimation, where negative examples are gathered from images of the scene with no people present. Each local model can be considered as a binary classifier for the metadata (e.g., head pose angle) associated with the training exemplar.

Calibrating the predictions of the local models provides for output values, which can be directly compared as confidence values of a query matching the local models. Calibration requires a separate training stage, using only labeled image patches containing heads. For each exemplar, positive examples are those patches for which the labels "match" and the remaining are negative. For example, in the case of head pan angle estimation, positive examples would correspond to image patches with pan angles within a specified threshold of the exemplar. As shown in Figure 5, calibration has the effect of dampening the output of less reliable detectors, while amplifying those that generalize better. Platt scaling is used to convert the raw SVM output of the post-calibration model to a probability value, which can be used as a detector confidence value. Figure 6 shows the top 5 exemplar detections for a query before and after calibration.

### 3.1.2 Single-Camera Algorithm

Let $\mathcal{D}$ represent the set of $M$ trained exemplar detectors, as described above. Each detector, $D_i$, is associated with the label, $y_i$, of the corresponding exemplar, and a scoring function $s_i(\cdot)$ that returns the Platt-scaled probability for a query example, $\mathbf{x_q}$. A query example, $\mathbf{x_q}$, corresponds to the feature representation for an image patch extracted from a roughly localized image window.



(a) Pre-calibration (b) Post-calibration

**Figure 5: For an ensemble of exemplars, calibration dampens the output of less reliable detectors, while amplifying those that generalize better.**


**Figure 6: For a query example (left), the top-scoring exemplars are shown before (top) and after (bottom) calibration.**

For each query in the search area, we obtain the top $K$ scoring detectors. The query with the best matches to head pose exemplars is retained as the head location prediction. To predict the head orientation at this location, we consider each of the top matching detectors as a noisy predictor of the query label, $y_q$, and model the ensemble prediction using a Mixture of Gaussians model. The contribution of each detector, $D_i$, is represented by a Gaussian with mean, $\mu = y_i$ and standard deviation, $\sigma = \frac{1}{\alpha s_i(\mathbf{x_q})}$, where $\alpha$ is a scaling parameter. Figure 7 shows an example of predicting the head pan angle of a query image using this approach.

## 3.2 Multi-Camera Head Pose Estimation

In a multi-camera network, predictions from multiple cameras can be aggregated by multiplying the resulting probability density functions from each camera. For clarity, all references to direction-based predictions are assumed to be from a global coordinate frame. Figure 8 shows an example of multi-camera prediction of head pan angle. This example is representative of the typical case where cameras that observe the front of the target's face provide more confident predictions. For the multi-camera system, the final prediction can be taken as the mode of the combined PDF.

The observation that certain viewpoints in a multi-camera setting are preferable motivates our approach for dynamic camera selection. Rather than aggregating the predictions from all the cameras in the network at once, we sequentially incorporate single-camera predictions until sufficient confidence is achieved. To estimate the number of cameras to sequentially sample to make a prediction, we incorporate the multi-class sequential probability ratio test [3]. We discretize the probability density function $p(y|\mathbf{x})$ of the per-camera prediction. For a discretized label, $y$, the ratio is

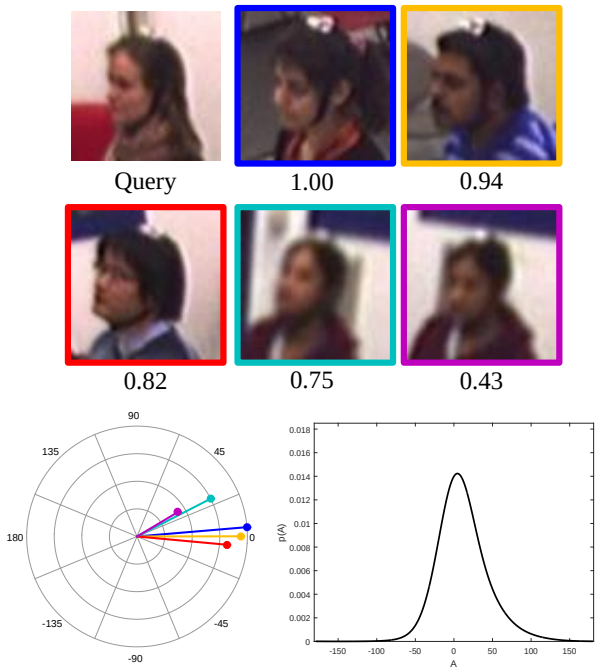



**Figure 7: A query image and the top scoring exemplars (top). The radial plot (bottom-left) shows the predicted pan angle and detector score for the corresponding exemplars. The pose angle predictions are combined into an ensemble estimate (bottom-right).**

defined as:

$$r(y|\mathbf{x}_{1:v}) = \frac{P(y|\mathbf{x}_{1:v})}{\sum_{y' \neq y} P(y'|\mathbf{x}_{1:v})} \quad (1)$$

where $\mathbf{x}_{1:v}$ denotes the input from a sequence of $v$ cameras. The class conditional probabilities, $P(y|\mathbf{x}_{1:v})$, are estimated using the Naive Bayes and uniform priors assumptions:

$$P(y|\mathbf{x}_{1:v}) = P(y|\mathbf{x}_{1:v-1})P(y|\mathbf{x}_v) \quad (2)$$

A prediction is made for a class when the ratio is greater than a user-specified threshold, $\tau$. A ratio greater than 1 indicates that the probability for a particular class is greater than the sum of the other choices.

## 3.3 Method Summary

Our method generalizes an efficient single-camera algorithm for head pose estimation to the multi-camera setting. It is applicable to a wide variety of multi-camera configurations, and, with dynamic camera selection, the computational efficiency does not necessarily grow with the number of cameras in the network. In the next section, we evaluate our approach on a benchmark dataset.

# 4. RESULTS

We evaluate our method on DPOSE [7], a publicly-available dataset for multi-camera head pose estimation, consisting of over 50,000 frames of 16 moving people captured by 4 calibrated, synchronized cameras. Figure 9 shows example frames from DPOSE, with a zoomed-in crop of the localized head region.

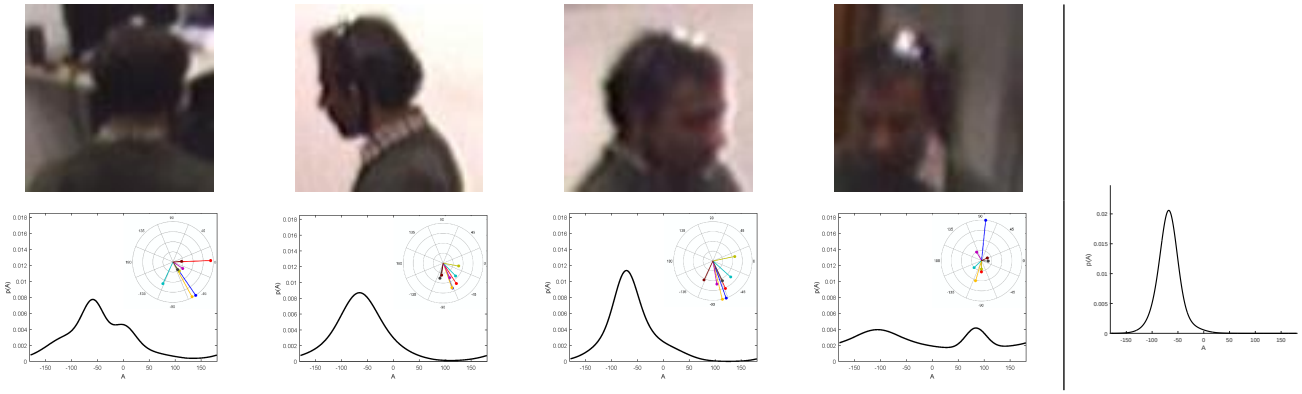


**Figure 8: For the query image from each view (top row), the top-scoring exemplar estimates are used to compute probabilistic predictions (bottom row), which are combined to give the system prediction (right).**

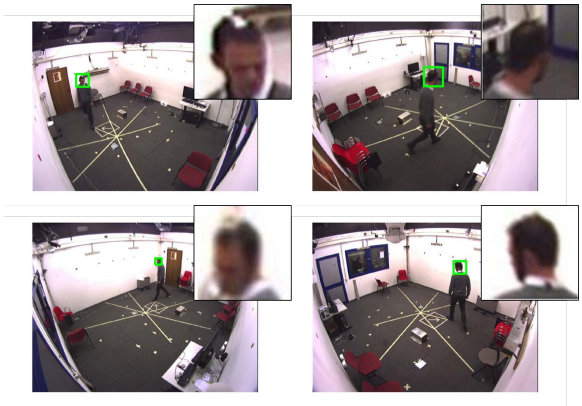


**Figure 9: DPOSE consists of labeled images from multiple people observed by four cameras.**

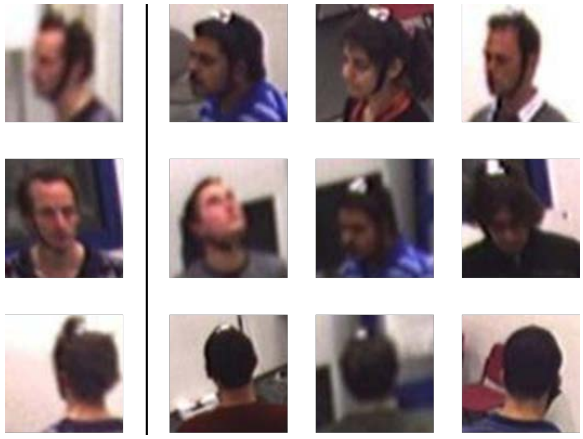


**Figure 10: For the image patch from each camera (column 1), the top scoring exemplars are shown.**

## 4.1 Exemplar Head Pose Learning

Targets are tracked using a multi-camera tracking algorithm that estimates a 3D bounding cube for each target [13]. The initial head search area and window size is based on the projected size of the target in a camera. From each rough localization image patch, square image patches are extracted and scaled to 70 x 70 pixels and represented using HOG features [2] with 7x7 cells and the 31-dimensional descriptor of Felzenszwalb et al. [5]. For ESVM learning, training and validation examples are randomly selected from DPOSE. For each training example, an exemplar model is trained. Negative examples are extracted from background images of the scene known not to contain people. Each exemplar model is calibrated using the validation examples. For head pose angle estimation, examples where the angle difference between exemplar and validation example is less than 10 degrees are taken as positives, and those greater than 90 degrees are negatives. Figure 10 shows the top matching exemplars for sample query examples from DPOSE.

## 4.2 Head Pose Estimation

Our approach provides real-valued predictions for both pan and tilt angles. To the best of our knowledge, no previous work has reported real-valued predictions on the DPOSE dataset for the problem of head pose estimation in multi-

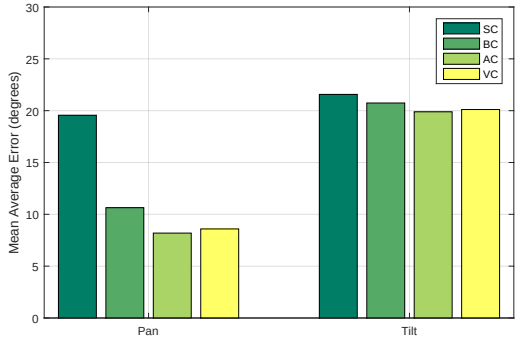



**Figure 11: Mean absolute error for head pan and tilt angle estimation.**

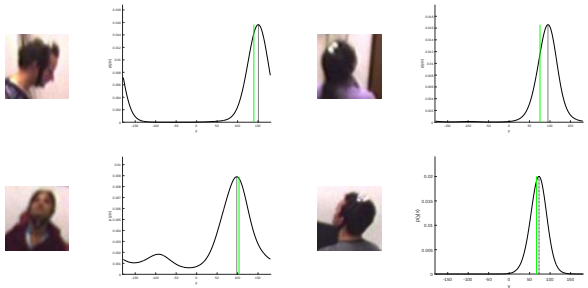


**Figure 12: Head pan angle predictions from our VC method for selected DPOSE image patches. Ground truth is shown in green, VC estimate in black.**

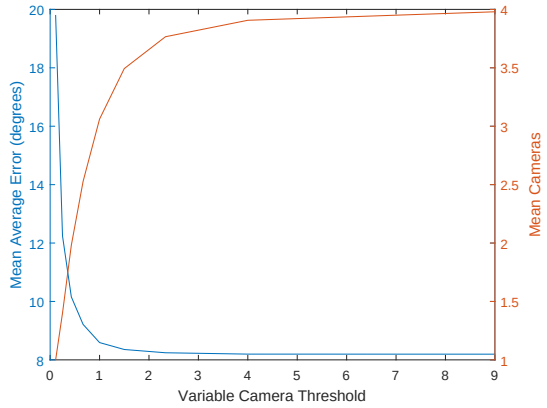



**Figure 13: For variable-camera selection, pose estimation error decreases and the number of cameras sampled increases as the threshold increases.**

camera networks. Here, we compare several variants of our method:

- *Single-camera (SC)* is the baseline approach, evaluated per-camera.
- *Best-camera (BC)* applies SC to each camera and returns the highest-confidence estimate.
- *All-cameras (AC)* applies SC to each camera and aggregates the predictions.
- *Variable-cameras (VC)* incorporates our dynamic camera selection scheme.

Each method used the same ensemble of exemplars. Two sets of head pose patches were used to train the models, 960 examples each for training and validation. The scaling constant, $\alpha = 0.05$ and the number of top-scoring exemplars, $K = 25$. In practice, our algorithm is robust to a range of values for these parameters. Figure 11 shows the results for head pose localization on DPOSE reported as the mean absolute error of the prediction compared to the provided ground truth over 1000 testing examples, averaged over 5 trials. The multi-camera methods (AC and VC) outperformed the single-camera methods (SC and BC), with the dynamic camera selection scheme (VC) performing comparably to the all-cameras (AC), with less computation. Figure 12 shows example predictions from our VC method.

For the variable-cameras (VC) method, the confidence threshold serves as a tunable parameter that changes the behavior from the single-camera to all-cameras paradigms. As such, pose estimation error decreases and the number of cameras sampled increases as the confidence threshold increases. Figure 13 shows these trends for head pan angle estimation with DPOSE. In our experiments, we set the variable camera threshold, $\tau = 1.0$.

**Table 1: Discrete head pose classification accuracy.**

| Method | Accuracy |
|---|---|
| Ours | 85.22% |
| Yan 2013 | 86.10% |
| Yan 2014 (HOG) | 80.00% |
| Yan 2014 (HOG+KL) | 87.00% |

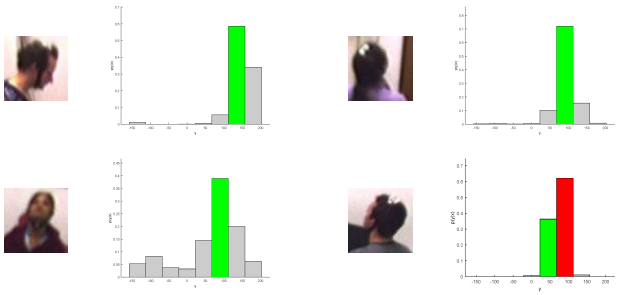


**Figure 14: For each image, the ground truth class is green and an incorrect classification is red.**

## 4.3 Discrete Head Pose Classification

Previous methods that have used DPOSE have only provided predictions for head pan angle into one of eight 45° bins. To compare our results to recent related work, we follow the ensemble exemplar learning approach previously described with the modification that the label associated with each exemplar corresponds to one of 8 classes rather than the provided real-valued ground truth. We follow the same experimental protocol as other recent work [17]. For training, the scene is divided into four quadrants and 30 training examples are randomly selected from each region for each of 8 quantized head poses. Results are averaged over 5 trials.

While our method was designed to provide precise real-valued predictions of head pose, it is competitive with the state of the art for the discrete classification task. Table 1 compares the classification accuracy of our method with several recently published approaches. Figure 14 shows example results from our method for this discrete prediction task. Closer inspection of the results shows that most of the errors in our variable-camera approach tend to lie within one discrete bin of the true pose. Since our method on this data achieves a mean absolute error of 8.59 degrees, it is likely that some portion of the misclassifications are due to quantization artifacts at the boundaries of the artificially-defined classes. Figure 15 shows the confusion matrix for the variable-camera pan classification experiment. Each row represents the true pose angle, and each column the angle predicted by our method. The diagonal banding illustrates the tendency of errors to fall in neighboring classes.

# 5. CONCLUSIONS AND FUTURE WORK

In this paper, we described a novel approach for head

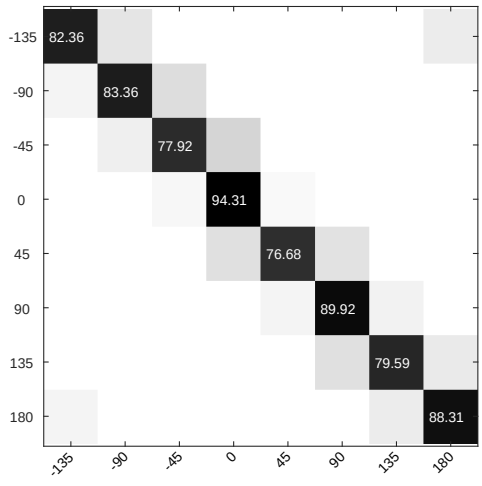



**Figure 15: Confusion matrix for discrete pan angle classification on DPOSE.**

pose estimation designed for multi-camera networks. Our framework is robust to low-resolution images, poorly localized bounding boxes, and appearance changes induced by changing person location in the scene. The computational requirements are also modest due to the use of inexpensive features and fast linear classifiers. In addition, we described a variable-camera scheme to dynamically select a subset of the available cameras for pose estimation, allowing for explicit trade-off between efficiency and accuracy. Experiments on a benchmark dataset show that our approach provides discrete classification accuracy on par with the state-of-the-art.

For future work, we plan to explicitly incorporate temporal smoothness in a tracking framework to better inform the camera selection process and reduce the overall number of views required for accurate estimation. Additionally, we plan to investigate methods to improve the single-camera prediction, with an eye toward further reducing computation requirements.

# 6. REFERENCES


[1] C. Canton-Ferrer, J. R. Casas, and M. Pardàs. Head pose detection based on fusion of multiple viewpoint information. In *Multimodal Technologies for Perception of Humans*, pages 305–310. Springer, 2007.

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

[3] J. W. Davis and A. Tyagi. Minimal-latency human action recognition using reliable-inference. *Image and Vision Computing*, 24(5):455–472, 2006.

[4] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.

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

[6] D. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 2010.

[7] A. K. Rajagopal, R. Subramanian, R. Vieriu, E. Ricci, O. Lanz, K. Ramakrishnan, and N. Sebe. An adaptation framework for head-pose classification in dynamic multi-view scenarios. In *Asian Conference on Computer Vision*, pages 652–666, 2013.

[8] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *Proc. International Conference on Computer Vision*, pages 89–96. IEEE, 2011.

[9] R. Munoz-Salinas, E. Yeguas-Bolivar, A. Saffiotti, and R. Medina-Carnicer. Multi-camera head pose estimation. *Machine Vision and Applications*, 23(3):479–490, 2012.

[10] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[11] J. Orozco, S. Gong, and T. Xiang. Head pose classification in crowded scenes. In *Proceedings of the British Machine Vision Conference*, volume 1, page 3, 2009.

[12] A. K. Rajagopal, R. Subramanian, E. Ricci, R. L. Vieriu, O. Lanz, K. Ramakrishnan, and N. Sebe. Exploring transfer learning approaches for head pose classification from multi-view surveillance images. *International Journal of Computer Vision*, 109(1-2):146–167, 2014.

[13] S. Spurlock and R. Souvenir. Pedestrian verification for multi-camera detection. In *Asian Conference on Computer Vision*, 2014.

[14] D. Tosato, M. Farenzena, M. Spera, V. Murino, and M. Cristani. Multi-class classification on riemannian manifolds for video surveillance. In *Proc. European Conference on Computer Vision*, pages 378–391. Springer, 2010.

[15] A. A. X. Zabulis, T. Sarmis. 3d head pose estimation from multiple distant views. *Proceedings of the British Machine Vision Conference*, 2009.

[16] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *Proc. International Conference on Computer Vision*, pages 1177–1184. IEEE, 2013.

[17] Y. Yan, R. Subramanian, E. Ricci, O. Lanz, and N. Sebe. Evaluating multi-task learning for multi-view head-pose classification in interactive environments. In *Proc. International Conference on Pattern Recognition*, pages 4182–4187. IEEE, 2014.

[18] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.