Discriminative Poses for Early Recognition in Multi-Camera Networks

Scott Spurlock, Junjie Shan, and Richard Souvenir
Department of Computer Science
University of North Carolina at Charlotte
9201 University City Blvd., Charlotte, NC 28223
{sspurloc, jshan1, souvenir}@uncc.edu

ABSTRACT

We present a framework for early action recognition in a multi-camera network. Our approach balances recognition accuracy with speed by dynamically selecting the best camera for classification. We follow an iterative clustering approach to learn sets of keyposes that are discriminative for recognition as well as for predicting the best camera for classification of future frames. Experiments on multi-camera datasets demonstrate the applicability of our view-shifting framework to the problem of early recognition.

CCS Concepts

•**Computing methodologies** → **Activity recognition and understanding;** *Supervised learning by classification;*

Keywords

early recognition; exemplar-based learning

1. INTRODUCTION

For human action recognition, certain poses are highly predictive for particular actions. Methods based on this observation have been applied in the single-camera setting [9, 14]. However, for distributed camera networks, most methods focus on more computationally expensive approaches, such as integrating multiple cameras or learning new 3D features. To take advantage of multiple viewpoints while retaining the computational efficiency of single-camera 2D recognition approaches, we revisit the idea of discriminative poses and adapt it to the multi-camera setting.

We present an approach to multi-camera action recognition that, for a given subject, dynamically selects the camera most likely to observe a *class-discriminative* pose. Figure 1 shows an example where some cameras in a network observe a class-discriminative pose sooner than others. Our method facilitates early recognition by learning *shift-discriminative* poses, which are not predictive for a particular class, but for a class-discriminative pose being observed *from a different camera in the network*. Our approach limits computation in two ways. First, instead of aggregating information from all of the cameras, one camera is dynamically selected for recognition at each frame. Second, our approach is applicable to early recognition and can predict an observed action prior to its completion. The contributions of this paper are: (1) iterative discriminative learning to identify keyposes; (2) dynamic view-shifting in multi-camera networks; and (3) discriminative poses for early recognition.

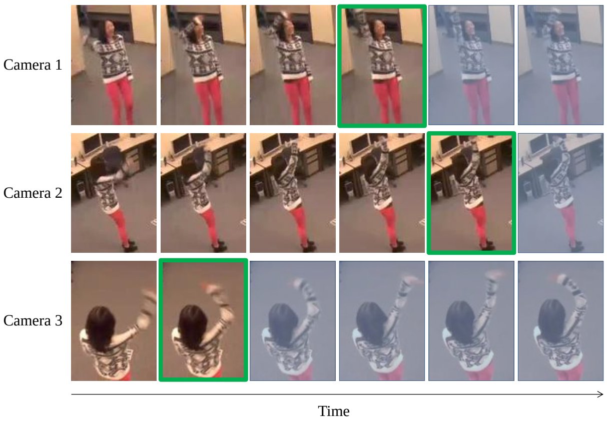


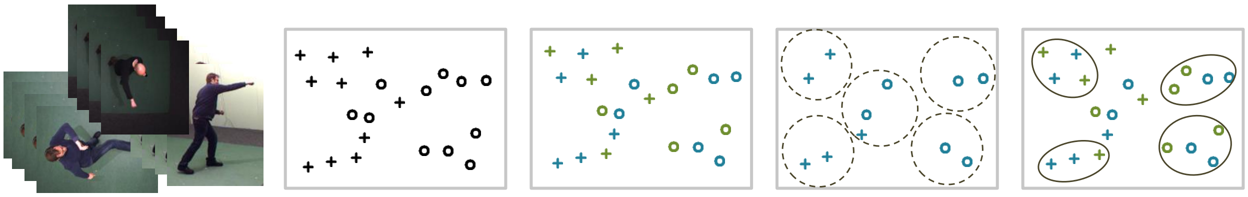
Figure 1: For this "wave", a class-discriminative pose (green box) is observed at different times in each camera. We present a method for early action recognition that dynamically selects the camera most likely to observe a discriminative pose.

2. RELATED WORK

There has been extensive work in human action recognition from video [20]. Many multi-camera methods use multi-view geometry to construct 3D models and solve 4D (3D plus time) recognition problems (e.g., [18]). These computationally-expensive methods require either 3D model construction or searching a large parameter space. Other approaches, rather than explicitly constructing 3D models, use a set of 2D image views with some aggregation scheme (e.g., [22]). These approaches tend to be computationally more efficient than the 4D methods, but still require computation for each camera in the network. Our approach neither fits 3D models nor requires computation at each camera per target, but selects a single view for recognition dynamically.

Other work has also considered the idea of selecting a sin-

ICDSC '15, September 08 - 11, 2015, Seville, Spain
© 2015 ACM. ISBN 978-1-4503-3681-9/15/09...$15.00
DOI: http://dx.doi.org/10.1145/2789116.2789117



Training Data 1. Training Examples 2. Split Data 3. Initialize Clusters 4. Refine Clusters

Figure 2: To learn a set of discriminative clusters of poses, classifiers are trained on local regions of feature space using cross-validation to prevent overfitting.

gle best view. Rudoy et al. [12] presents a method to identify the best viewpoint for an action from a human observer's standpoint based on motion features. Other approaches select the best view based on the number of spatio-temporal features detected [21], estimated camera distances and person orientation [15], or silhouette properties [10]. Unlike our approach, which learns to predict the next best view based solely on the current view, these methods still require features to be computed for every camera at each time step.

Several approaches have sought to learn keyposes for action recognition. Cheema et al. [1] learns weights based on the predicted discrimination power of poses learned in an unsupervised fashion. By contrast, other methods (e.g., [23]) directly incorporate measures of discriminative power to select keyposes. Methods based on Exemplar SVM [11] have been used to learn discriminative features for action recognition [6] or visual words for scene recognition [3].

There has been some recent work in early event detection from video. One method [2] used reliable-inference to predict the beginning of event, but required manual configuration to fit the data. Schindler and Van Gool [14] observed that short sequences are often sufficient to recognize a longer sequence. Another approach computes bags of features for each possible action subsequence and matches new observations probabilistically with dynamic programming [13], but requires a priori specification of the fraction of the sequence to be observed. Max-Margin Event Detection (MMED) incorporates a variant of structured output SVM to detect incomplete actions [5].

Our approach extends the ideas of discriminative features and early event detection to the multi-class setting resulting in an algorithm that achieves performance on par with more computationally expensive action recognition approaches.

3. DISCRIMINATIVE POSES

Discriminative feature learning incorporates class information during the dictionary learning phase of training to learn a set of features that are more discriminative during the later stages of classification than features obtained in an unsupervised fashion (e.g., k-means clustering). We apply this concept to learn both class- and shift-discriminative poses for the problem of action recognition in multi-camera settings.

3.1 Discriminative Pose Learning

Let $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_C\}$ be a set of synchronized image streams from C different cameras. Each stream, $\mathbf{X}_c = \{\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, \ldots\}$, could include example actions from a variety of actors and relative positions. Each input sample, $\mathbf{x}_{c,i} \in \mathcal{R}^D$, corresponds to a D-dimensional feature representation, which could be any descriptor used for action recognition (e.g., [8, 17]).[1] Let $\mathbf{y} = \{y_1, y_2, \ldots\}$ represent the associated class labels for each frame.

We follow a recent approach for iterative discriminative cluster learning, where, for each class, the goal is to iteratively learn an ensemble of discriminative local support vector machine (SVM) classifiers [16]. Figure 2 illustrates the method. For each class, the training examples are divided into training and validation sets. For any positive training example in a homogenous region (i.e., the ratio of positive to total neighbors is above a threshold), a classifier is trained using neighboring positive examples and all the negative examples in the set. The classifier is applied to the validation set, and the highest-scoring correct positive examples are retained for training the classifier in the next round. The classifiers are iteratively re-trained, each time swapping the training and validation sets until the cluster membership converges. In practice, the process converges quickly, typically requiring less than five iterations.

To normalize the responses of the local SVM clusters, the SVM scores are converted to posterior probabilities using Platt scaling. Let $\mathcal{C}$ represent the ensemble of SVM models returned by iterative discriminative cluster learning. For an example, $\mathbf{x}$, and label, y, the posterior probability of $\mathbf{x}$, $P(y|\mathcal{C}, \mathbf{x})$ is defined to be the maximum posterior probability of the cluster SVMs in $\mathcal{C}$ associated with class label y. Let $\phi(\mathcal{C}, \mathbf{x})$ be a membership function that returns the label for example $\mathbf{x}$ using discriminative cluster set, $\mathcal{C}$:

$$\phi(\mathcal{C}, \mathbf{x}) = \begin{cases} \operatorname*{argmax}_y P(y|\mathcal{C}, \mathbf{x}) & \text{if } \max P(y|\mathcal{C}, \mathbf{x}) > \tau \\ \varnothing & \text{otherwise} \end{cases} \tag{1}$$

where τ is a classification threshold. In our implementation, we learn this value by cross-validation.

3.2 Class-Discriminative Poses

Class-discriminative poses can be obtained by direct application of discriminative pose learning. For a given training set, $\langle \mathcal{X}, \mathbf{y} \rangle$, for this phase only the class label is relevant; the viewpoint and particular sequence that generated the example are not used. Figure 3 shows some of the poses identified as class-discriminative using this method.

3.3 Shift-Discriminative Poses

The intuition for shift-discriminative poses is that certain poses may not necessarily indicate that a *particular* action is taking place, but that *some* action may be taking place

[1] Depending on which is more natural in context, the terms "frame" and "pose" will be used interchangeably to mean "the feature representation of the subject in the frame".

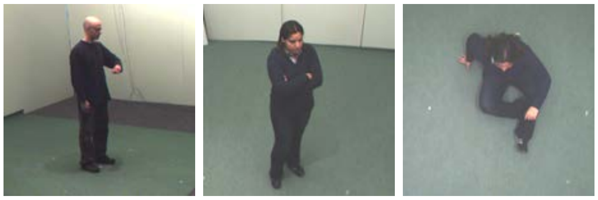


Figure 3: Class-discriminative poses (check watch, cross arms, sit).

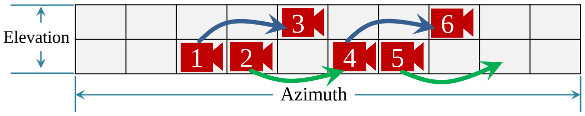


Figure 4: Each bin represents a discretized portion of the viewsphere; in this case, there are 10 azimuth and 2 elevation bins. The icons represent the relative locations of cameras. Arrows of the same color indicate the same (relative) view-shift.

and it may be preferable to observe the motion from a different viewpoint. That is, rather than being indicative of a particular class, shift-discriminative poses are indicative of a *view-shift*, specifically a shift to another view in the camera network more likely to observe a class-discriminative pose.

A view-shift is a relative change of viewpoint. Potential view-shifts are determined by the physical location of the cameras relative to the target. The view half-sphere is discretized into a fixed number of azimuth and elevation offsets.[2] Let c_m and c_n be the discretized location of two cameras in the network, represented in (cyclic) azimuth and elevation; the view-shift between them is represented as $\vec{v} = c_n - c_m$. $\langle 0, 0\rangle$ represents maintaining the current view, while $\langle +1, -1\rangle$, represents shifting to a camera one offset around (azimuth) and one offset up (elevation) the view-sphere. Figure 4 shows an example configuration where the view-sphere has been divided into two elevation and ten azimuth bins. For $C = 6$ cameras, there are at most 30 $(C \times (C-1))$ possible view-shifts. The actual number of potential view-shifts is often lower as there are shared view-shifts. For example, in Figure 4, the view-shift $\langle +2, -1\rangle$ describes both the shift from camera 1 to 3 and 4 to 6.

For learning shift-discriminative poses, the viewpoint and sequence membership of each frame are maintained during training. Figure 5 represents a network with 5 cameras; each row represents a sequence from a particular camera, each unit represents a frame of video, and green boxes represent class-discriminative poses. Cameras 1, 3, and 4 will encounter class-discriminative poses, while cameras 2 and 5 will not. Our goal is to learn when to view-shift to maximize the likelihood of observing class-discriminative poses.

Shift-discriminative learning occurs in two stages. Based on the set of class-discriminative clusters, $\mathcal{C}_C$, training examples are assigned (potentially multiple) view-shift labels

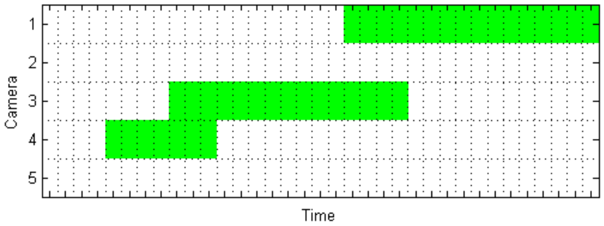


Figure 5: For a sequence captured by five cameras, class-discriminative poses are shown in green.

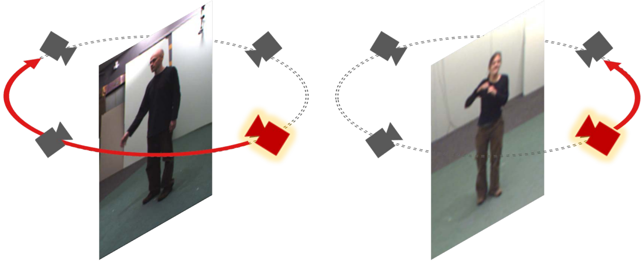


Figure 6: Examples of shift-discriminative poses from the IXMAS data set. The arrows illustrate associated view-shifts to alternate cameras.

by the following recursively-defined rule:

$$\Psi(\mathbf{x}_{c,i}, \vec{v}) = \begin{cases} 0 & \text{if } x_{c',i'} = \varnothing \\ 1 & \text{if } \phi(\mathcal{C}_C, \mathbf{x}_{c',i'}) = y_i \\ -1 & \text{if } \phi(\mathcal{C}_C, \mathbf{x}_{c',i'}) \neq \varnothing \\ \Psi\left(\mathbf{x}_{j,i'}, \langle 0,0\rangle\right) & \text{otherwise} \end{cases} \qquad (2)$$

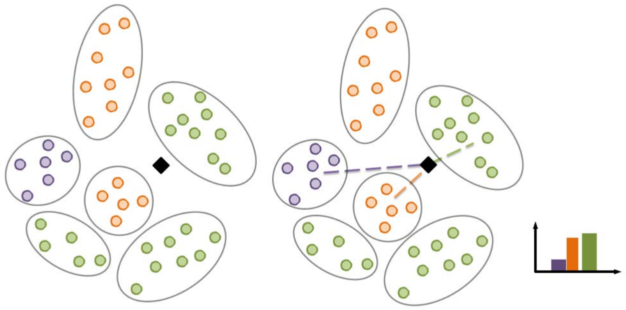
where Ψ is a compatibility function between frame $\mathbf{x}_{j,i}$ and view-shift $\vec{v}$, $c' = c + \vec{v}$ represents the resulting camera location after applying view-shift, $\vec{v}$, and i' represents the index of the next frame in the training sequence. In short, for a given pose, $\mathbf{x}_{c,i}$, if a particular view-shift, $\vec{v}$, leads to a correct class-discriminative pose in the training data, the pose represented by $\mathbf{x}_{c,i}$ is considered a positive example for view-shift, $\vec{v}$. If, however, view-shift, $\vec{v}$, leads to an *incorrect* class-discriminative pose (i.e., a training pose misidentified by a class-discriminative detector), the pose represented by $\mathbf{x}_{c,i}$ is considered a negative example for view-shift, $\vec{v}$.

The next stage extends discriminative pose learning to the multi-label case. For each possible view-shift, $\vec{v}$, each pose is positive, negative, or neutral. Let $\mathcal{X}_{\vec{v}}$ represent the subset of training examples that are either positive or negative with respect to view-shift, $\vec{v}$. We apply the discriminative pose learning algorithm from Section 3.1 to the set $\mathcal{X}_{\vec{v}}$ for each possible view-shift. The result is a set of shift-discriminant poses. Figure 6 shows two examples, where the arrow shows the relative shift associated with the keypose.

4. RECOGNITION ALGORITHMS

Previous work employing discriminative features most often considers hard cluster assignments where a particular feature is discriminative or not. Viewed in a different light, the set of local discriminative classifiers for a given class serve to define a complex decision boundary in feature space that could be used for soft assignment. Figure 7 depicts this

[2] Most action recognition features are invariant to small shifts and scale. This allows changes in viewpoint relative to the subject to be represented by 2D translations on the viewsphere.



(a) Hard assignment (b) Soft assignment

Figure 7: For hard assignments (a), the new pose (black square) is not considered discriminative. For soft assignments (b), the class posterior distribution for the new pose is based on the distances to the nearest discriminative clusters.

situation for a toy example in a 2D feature space.

These two viewpoints on pose classification directly inform the two variants of our action recognition algorithm, which we call exemplar recognition and sequential recognition. In both cases, frames are processed by a single camera at a time with no data buffering.

4.1 Exemplar Recognition

This version uses hard assignments. When a shift-discriminative pose is observed, a view-shift is applied. When a class-discriminative pose is observed, a prediction is made, which ends processing for the sequence.

```
Algorithm 1: Exemplar Recognition
Input: Synchronized pose sequences, $\mathcal{X}$; discriminative
       clusters, $\mathcal{C}_C, \mathcal{C}_S$; camera locations, $\{c_m\}$
Output: Predicted label, $\hat{y}$
$a \equiv$ index of active camera
Current time, $t \leftarrow 1$
while not yet classified do
    $\hat{y} \leftarrow \phi(\mathcal{C}_C, \mathbf{x}_{a,t})$
    if $\hat{y} \neq \varnothing$ then
        return $\hat{y}$
    (Possible) view-shift: $a \leftarrow a + \phi(\mathcal{C}_S, \mathbf{x}_{a,t})$
    $t \leftarrow t + 1$
```

4.2 Sequential Recognition

This version incorporates soft assignments based on the class posteriors. Rather than making a prediction once a single discriminative pose is observed, we incorporate the multi-class sequential probability ratio test proposed by Davis and Tyagi [2]. For a sequence of T observations, the ratio is defined as:

$$r(y|\mathbf{x}_{1:T}) = \frac{P(y|\mathbf{x}_{1:T})}{\sum_{y' \neq y} P(y'|\mathbf{x}_{1:T})}. \tag{3}$$

Class conditional probabilities, $P(y|\mathbf{x}_{1:T})$, are estimated using the Naive Bayes and uniform priors assumptions:

$$P(\mathbf{y}|\mathbf{x}_{1:T}) = P(\mathbf{y}|\mathbf{x}_{1:T-1})P(\mathbf{y}|\mathcal{C}, \mathbf{x}_T) \tag{4}$$

A prediction is made when, for particular class, y, the ratio, r, is greater than a pre-determined threshold. $r(y|\mathbf{x}_{1:T}) > 1$ indicates that the probability for a particular class is greater than the sum of the other choices.

```
Algorithm 2: Sequential Recognition
Input: Synchronized pose sequences, $\mathcal{X}$; discriminative
       clusters, $\mathcal{C}_C, \mathcal{C}_S$; camera locations, $\{c_m\}$
Output: Predicted label, $\hat{y}$
$a \equiv$ index of active camera
Current time, $t \leftarrow 1$
Initialize $P(\mathbf{y}|\mathbf{x}_{0:0})$ to uniform distribution
while not yet classified do
    Update $P(\mathbf{y}|\mathbf{x}_{1:t})$ (Eq. 4)
    if $\exists y', r(y'|\mathbf{x}_{1:t}) > \tau$ then
        return $y'$
    Update $P(\vec{v}|\mathbf{x}_{1:t})$ (Eq. 4)
    if $\exists \vec{v}', r(\vec{v}'|\mathbf{x}_{1:t}) > \tau$ then
        view-shift: $a \leftarrow a + \phi(\mathcal{C}_S, \mathbf{x}_{a,t})$
    $t \leftarrow t + 1$
```

5. EXPERIMENTS

We evaluate our two algorithms, VS_{EXM} (exemplar recognition) and VS_{SEQ} (sequential recognition), on two multi-view human action recognition datasets: i3DPost [4] (8 actors, 10 actions) and INRIA Xmas Motion Acquisition Sequences (IXMAS) [19] (10 actors, 11 actions).

View-shifting can be applied with any frame-based feature descriptor; we use the Motion Context descriptor [17], which represents the distribution of occupancy and $x-$ and $y-$ components of optic flow in a subject's bounding box. For discriminative cluster learning, we followed the guidance in [16]. The cluster initialization threshold is 40% class label purity for $k = 20$ nearest neighbors. For cluster SVM, the positive example threshold is -0.9. During training, each discriminative SVM maintains between 2 and 5 positive examples. The detection thresholds, τ_c and τ_s, are determined using cross-validation.

5.1 Pose Discrimination

For dictionary learning, we compare discriminative clustering to k-means (KM), commonly used for unsupervised dictionary learning, and Submodular Dictionary Learning (SDL) [7], a supervised dictionary learning method that optimizes cluster compactness, element similarity, and class discriminativeness. For k-means and SDL, $k = 1000$ and the pose-level assignment is based on the majority class of the cluster the pose is assigned. Using IXMAS, 7 actors were used for training and 3 for testing. Table 1 shows the results of the frame-level classification accuracy.

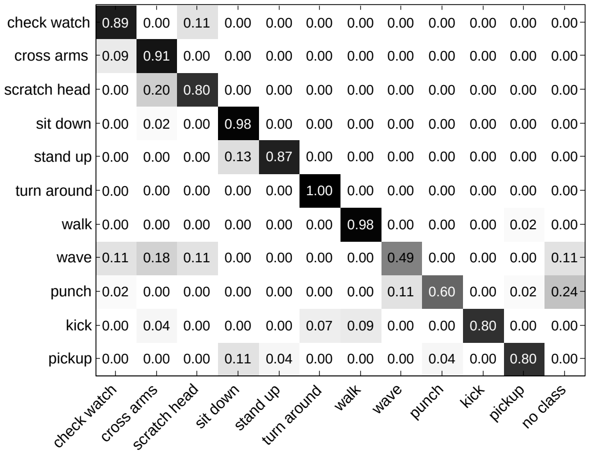
The EXM variant shows high precision and low recall since poses from heterogeneous regions of feature space are excluded from training discriminative clusters, and, unlike the

Table 1: Precision and recall on ~13,000 labeled frames from IXMAS.

	EXM	SEQ	k-means	SDL
Precision	**0.92**	0.49	0.40	0.35
Recall	0.09	**0.47**	0.45	0.40

Table 2: Recognition accuracy using discriminative poses and various multi-camera prediction schemes.

	Exemplar			Sequential		
	SC	MC	VS	SC	MC	VS
IXMAS	0.69	0.67	**0.77**	0.69	**0.84**	0.83
i3DPost	0.60	0.63	**0.75**	0.62	**0.77**	0.70



	check watch	cross arms	scratch head	sit down	stand up	turn around	walk	wave	punch	kick	pickup	no class
check watch	0.89	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
cross arms	0.09	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
scratch head	0.00	0.20	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sit down	0.00	0.02	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
stand up	0.00	0.00	0.00	0.13	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00
turn around	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
walk	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.02	0.00
wave	0.11	0.18	0.11	0.00	0.00	0.00	0.00	0.49	0.00	0.00	0.00	0.11
punch	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.60	0.00	0.02	0.24
kick	0.00	0.04	0.00	0.00	0.00	0.07	0.09	0.00	0.00	0.80	0.00	0.00
pickup	0.00	0.00	0.00	0.11	0.04	0.00	0.00	0.00	0.04	0.00	0.80	0.00

Figure 8: Confusion matrix for VS_{SEQ} on IXMAS for early recognition. On average, actions were identified after observing 24% of the action.)

other methods, does not make a prediction for most input poses. The SEQ method has higher recall and precision than both k-means and SDL and the highest value for recall among all the methods. Our approaches learn highly discriminative features, which can be used to improve sequence-scale classification.

5.2 View-Shifting

We implemented two baseline approaches to evaluate view-shifting. The single-camera (SC) method operates from a single viewpoint without view-shifting, and the multi-camera (MC) method aggregates information from all cameras. Both methods can be evaluated using the exemplar (EXM) and sequential (SEQ) prediction schemes. For MC_{EXM}, a sequence is classified when the first class-discriminative pose is observed in any camera. For MC_{SEQ}, a single, combined posterior is calculated based on the observed poses from all cameras; a sequence is classified based on the ratio test described in Section 3.

We measured performance for each dataset following the experimental protocols most commonly found in the literature. For IXMAS, this is leave-one-actor-out (LOAO) with the results averaged across all the iterations. For i3DPost, the first 5 actors are used for the training, and the last 3 for testing. Table 2 shows the recognition accuracy.

For both datasets, the view-shifting methods out-perform the single-camera baselines and give similar or better accuracy than the multi-camera approaches. This performance comes at a fraction of the computational cost, since the VS methods only process frames from a single camera at a time. In the exemplar (hard assignment) scheme, the view-shift approach outperforms both SC and MC by a wide margin.

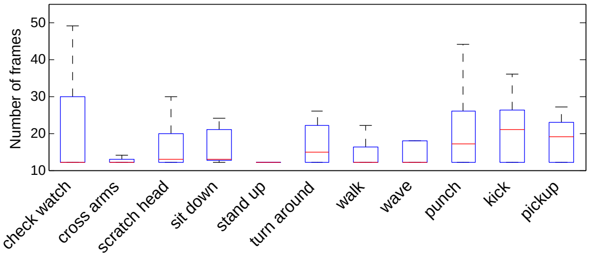


Figure 9: The box plots show the number of frames observed for each action by VS_{SEQ} prior to a prediction during testing on IXMAS data.

Table 3: Early recognition accuracy and percent of sequence observed on benchmark datasets.

	IXMAS		i3DPost	
	Acc.	% Seq.	Acc.	% Seq.
VS_{EXM}	0.77	0.21	**0.75**	0.25
VS_{SEQ}	**0.83**	0.24	0.70	0.27
SOSVM-SC	0.52	0.20	0.52	0.41
SOSVM-MC	0.65	0.15	0.60	0.32

Across datasets and camera schemes, sequential variants outperformed exemplar variants. The sequential approaches accumulate evidence for a prediction over multiple frames compared to the exemplar approaches. Figure 8 shows the confusion matrix for VS_{SEQ} on IXMAS data. Many of the actions are recognized at or near 100% (e.g., sit, turn, walk). The wave action was confused with other similar-starting actions (e.g., check watch, cross arms, scratch head). Figure 9 shows the number of frames observed prior to classification. In general, there does not appear to be a correlation between accuracy and number of frames observed. For example, the accuracy for the check watch and cross arms actions is similar, but the number of frames observed varies significantly.

5.3 Early Action Recognition

We implemented an approach based on structured output SVM (SOSVM), as described in [5]. For SOSVM, the frame-level features are aggregated into histograms of quantized words to represent video sequences. For each dataset, the dictionary size, k, was selected to produce the best results, $k = 1100$ for IXMAS and $k = 500$ for i3DPost. We trained a separate detector for each class and adapted the classifiers to the multi-camera setting. SOSVM-SC operates on single-camera input, predicting the class of the first detector to fire while observing a sequence. SOSVM-MC predicts the class of the first detector to fire for any camera in the network. Table 3 shows the recognition accuracy and average fraction of frames observed for each of the experiments.

Both view-shifting approaches outperformed the SOSVM detector. Across methods, the average fraction of the sequence observed was not a strong predictor of the final classification accuracy. However, for each method, a sequence was more likely to be classified correctly when classification was delayed. The plots in Figure 10 show precision and recall as a function of the fraction of the sequence observed for each of the methods on IXMAS.

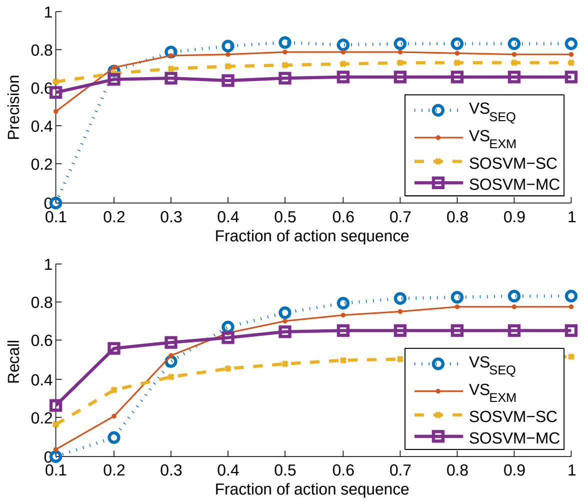


Figure 10: Precision and recall plots as a function of the fraction of the action observed for early event recognition on IXMAS.

6. CONCLUSIONS AND FUTURE WORK

This paper focused on the problem of early action recognition in multi-camera networks. We proposed a discriminative keypose-based approach that achieved similar accuracy to approaches using all cameras, while processing a single camera at a time. Our method is applicable to a wide variety of image features and distributed camera network architectures. For the future, we plan to incorporate hierarchical cascade models for additional computational savings. We also plan to investigate transferability of learned pose models between different camera configurations.

7. REFERENCES

[1] S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage. Action recognition by learning discriminative key poses. In *IEEE Intl Conf. on Computer Vision Workshops*, pages 1302–1309, 2011.

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

[3] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Trans. Graph.*, 31(4):101, 2012.

[4] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas. The i3dpost multi-view and 3d human action/interaction database. In *Visual Media Production*, pages 159–168. IEEE, 2009.

[5] M. Hoai and F. De la Torre. Max-margin early event detectors. *Intl Journal of Computer Vision*, 107(2):191–202, 2014.

[6] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis. Representing videos using mid-level discriminative patches. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2571–2578. IEEE, 2013.

[7] Z. Jiang, G. Zhang, and L. S. Davis. Submodular dictionary learning for sparse coding. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3418–3425. IEEE, 2012.

[8] I. Laptev. On space-time interest points. *Intl Journal of Computer Vision*, 64(2-3):107–123, 2005.

[9] L. Liu, L. Shao, X. Zhen, and X. Li. Learning discriminative key poses for action recognition. *IEEE T. Cybernetics*, 43(6):1860–1870, 2013.

[10] T. Määttä, A. Härmä, and H. Aghajan. On efficient use of multi-view data for activity recognition. In *Proc. Intl Conf. on Distributed Smart Cameras*, ICDSC '10, pages 158–165, New York, NY, USA, 2010. ACM.

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

[12] D. Rudoy and L. Zelnik-Manor. Viewpoint selection for human actions. *Intl Journal of Computer Vision*, 97(3):243–254, 2012.

[13] M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proc. Intl Conf. on Computer Vision*, pages 1036–1043. IEEE, 2011.

[14] K. Schindler and L. Van Gool. Action snippets: How many frames does human action recognition require? In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[15] C. Shen, C. Zhang, and S. Fels. A multi-camera surveillance system that estimates quality-of-view measurement. In *Proc. Intl Conf. on Image Processing*, volume 3, pages III–193. IEEE, 2007.

[16] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Proc. European Conf. on Computer Vision*, pages 73–86. Springer, 2012.

[17] D. Tran and A. Sorokin. Human activity recognition with metric learning. In *Proc. European Conf. on Computer Vision*, pages 548–561. Springer-Verlag, 2008.

[18] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *Proc. Intl Conf. on Computer Vision*, pages 1–7, 2007.

[19] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257, 2006.

[20] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.

[21] C. Wu, A. H. Khalili, and H. Aghajan. Multiview activity recognition in smart homes with spatio-temporal features. In *Proc. Intl Conf. on Distributed Smart Cameras*, pages 142–149. ACM, 2010.

[22] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 489–496, 2011.

[23] Z. Zhao and A. M. Elgammal. Information theoretic key frame selection for action recognition. In *Proc. of the British Machine Vision Conf.*, pages 1–10, 2008.