Action Recognition in the Presence of One Egocentric and Multiple Static Cameras

Bilge Soran, Ali Farhadi, Linda Shapiro

University of Washington, Dept. of Computer Science and Engineering
{bilge, ali, shapiro}@cs.washington.edu

Abstract. In this paper, we study the problem of recognizing human actions in the presence of a single egocentric camera and multiple static cameras. Some actions are better presented in static cameras, where the whole body of an actor and the context of actions are visible. Some other actions are better recognized in egocentric cameras, where subtle movements of hands and complex object interactions are visible. In this paper, we introduce a model that can benefit from the best of both worlds by learning to predict the importance of each camera in recognizing actions in each frame. By joint discriminative learning of latent camera importance variables and action classifiers, our model achieves successful results in the challenging CMU-MMAC dataset. Our experimental results show significant gain in learning to use the cameras according to their predicted importance. The learned latent variables provide a level of understanding of a scene that enables automatic cinematography by smoothly switching between cameras in order to maximize the amount of relevant information in each frame.

1 Introduction

Activities that people perform in their daily lives span a wide spectrum of actions. Recognizing some actions requires reasoning about complex human-object interactions and detailed observation of the actions. For example, recognizing the cracking of an egg requires observations about the state change of the egg and characteristic postures of the hand. Some other actions, like walking to a refrigerator, are better recognized when a holistic view of an actor is visible. The movement of the human body provides strong cues for these kinds of activities.

The conventional setting of activity recognition involves studying the behavior of an actor from one or multiple static cameras [1]. There has been significant improvement over the last decade on recognizing actions that require observing the movements of the human body. However, in this setting, there are major challenges in recognizing actions that require subtle movements/gestures. This is mainly due to severe occlusions and distractions from image regions where the actual action is taking place.

An alternative is to use egocentric cameras (also called first-person or wearable cameras), with which the actions are observed from the actor's perspective ([2]). Although less susceptible to occlusion, egocentric cameras provide their

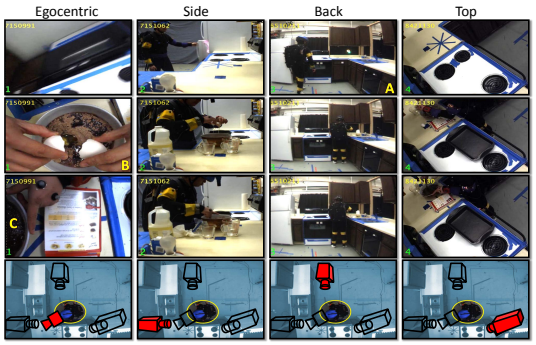

Fig. 1. We study action recognition in the presence of a single egocentric camera and multiple static cameras. The bottom row in Figure 1 illustrates settings where videos of people making brownies have been recorded from an egocentric camera (first column) and three static cameras (columns 2, 3, 4). Actions like cracking an egg (image B) are better recognized using an egocentric camera where information about subtle movements and complex interaction is available. However, information about holistic body movements is typically missing in egocentric cameras. Actions like walking are better recognized from a static camera (image A). Furthermore, people tend to look away when they perform actions that become procedural memories to them. For example the subject in image C looks at the recipe (instead of looking at the bowl) while stirring the brownie mix. This results in missing valuable action information in egocentric cameras. In this paper, we show a model that can benefit from both egocentric and static cameras by reasoning about the importance of each camera for each action.

own set of challenges. For example, the human body, which is one of the main cues for some actions, is not visible in an egocentric camera. The camera has complex motion resulting in frequent blurs and appearance distortions. Furthermore, people tend to look away when they perform actions they are comfortable with. This results in losing major parts of signals that correspond to the main action. For example, when stirring a food mixture, people look around to determine what ingredients they need next, check the time, or read the next step in the recipe. The image marked with "C" in Figure 1 corresponds to a sample frame from a stirring action in which the actor is actually reading the recipe.

Our goal in this paper is to study the problem of understanding human actions in the presence of a single egocentric and multiple static cameras. If an oracle provides information about the importance of each camera, then the problem of recognizing human actions becomes a multi-modal classification problem. In our formulation we use a latent variable to encode the importance of each camera for each frame and introduce a model to jointly learn the latent camera variables and the action classifiers.

Our experimental results show significant success on the challenging CMU-MMAC dataset that includes both static and egocentric cameras. As a side product, our method enables automatic cinematography. In the presence of an

egocentric and multiple static cameras, our method can automatically select a camera through which the action is better observed. By enforcing smooth transitions between cameras, we can automatically direct a scene with multiple cameras.

2 Related Work

Human activity recognition has attracted several researchers over the last decade. Comprehensive surveys are provided in [1, 3]. We categorize related work in multi-view action recognition, egocentric activity recognition and a brief summary of virtual cinematography.

Egocentric Action Recognition typically refers to studying human actions from a camera that is mounted on the body (head to chest). To address the challenges and characteristic constraints of egocentric action recognition, [2] uses a SIFT-based representation. [4] used source constrained clustering to classify actions from egocentric videos. Hand-object interactions are used by [5], where an object-based representation for action recognition that jointly models the objects and actions is proposed. A generative probabilistic model that recognizes actions while predicting gaze locations is proposed by [6]. Hand motion and gaze information for egocentric activity recognition is used by [7]. A novel activities-of-daily-living (ADL) dataset is provided by [8], on which the change in the appearance of the objects in interaction for recognizing activities is explicitly modeled. The problem of understanding simple social interactions in egocentric settings is studied by [9], while [10] examines action recognition in sports videos using egocentric cameras. [11] uses eye movements and ego-motions to better recognize indoor activities. [12] utilizes egocentric action recognition for the purpose of contextual mapping. [13] studies the problem of activity recognition using low resolution wearable cameras. The affect of gaze prediction in action recognition is explored by [14]. Actions can also be modeled through state transitions as suggested by [15]. From a different perspective, [16] handles a different action recognition problem: trying to understand the other person's interaction with the camera wearer with respect to an egocentric camera.

Multi-view Action Recognition: Multi-view action recognition has been approached by learning latent variables that encode the change in the appearance of actions or view points of actions [17–19], by joint learning of shared structures across multiple views [20], by hierarchical models of spatio-temporal features [21], by local partitioning and hierarchical classification of 3D HOG descriptors [22], by transfer learning [23–25] and by using spatio-temporal self-similarity descriptors [26]. Multiple datasets exists for multi-view action recognition: i3dpost [27], IXMAS [28, 22], and CMU-MMAC [29]. We use static and egocentric recordings of the CMU-MMAC dataset in this paper. [30, 31] have studied action recognition by combining egocentric cameras with IMU sensors using the CMU-MMAC dataset. [32, 33] used IMU sensors in the same dataset to recognize actions. To the best of our knowledge, there has been no study for activity recognition using an egocentric camera and multiple static cameras. Note that approaches that re-

quire 3D reconstruction of the subject or visual hull are not directly applicable to our settings that uses egocentric cameras.

Virtual Cinematography: The majority of the work has focused on active and interactive cinematography, where one has control over the position of the cameras and other parameters such as lighting conditions [34, 35]. We are mainly concerned with a passive case where multiple videos of a scene exist, and one needs to decide which camera to use for each frame. Virtual cinematography is not the main focus of this paper, and our model does not address the principles of cinematography. We merely show that our model provides a level of understanding that enables camera selection.

3 Approach

Our task is to recognize human actions in the presence of a single egocentric and multiple static cameras. Our intuition is that each camera plays a different role in predicting an action and should be weighted according to its importance. Given the importance of all cameras for recognizing the action in a frame, the problem of action recognition reduces to a multi-modal classification problem. Unfortunately, the camera importance information is not available. We jointly learn the importance of cameras along with the action classifiers.

During training we are given videos from a single egocentric and multiple static cameras and action labels for each frame. At test time, the task is to assign an action label to each frame given the observations from all the cameras. To set up notation, let us assume that there are C cameras, N frames, and M different actions. Frame i from camera j is represented by a d-dimensional feature vector x_i^j where $i \in \{1:N\}$, and $j \in \{1:C\}$. Each frame i is also labeled with the action label y_i where $y_i \in \{1:M\}$. The importance of each camera j in correctly understanding the action in frame i is represented by $\alpha_i^j \in [0,1]$. Our model aims at coupling the tasks of predicting action classifiers and camera importance variables.

Intuitively, the choice of α should depend on both the observations from all cameras and the action of interest. For example, the static side camera may be more informative than the egocentric camera in encoding walking, while for cracking an egg an egocentric camera is preferred. The importance of cameras may vary during actions. For example, at the beginning of an action like "taking a pan out of an oven" where the movement of the person toward the oven is informative, one might assign more weight to the side static camera. Toward the end of the action where the actor is reaching for the pan inside the oven, the egocentric camera becomes more important. Our model takes the importance of each camera into account while making predictions about actions.

The best estimate of the camera importance variables is the one that maximizes the accuracy of action prediction. We adopt a bi-linear multi-class max margin formulation where the importance of each camera is modeled by an element-wise product operator $\odot$ and a vector A of all latent camera importance variables. We stack all the observations across all cameras into an observation

vector $\mathcal{X}_i = [x_i^1, ..., x_i^C]$ which is a $(C * d)$-dimensional representation of frame i, where C is the total number of cameras and x_i^j is a d-dimensional vector of each frame i in camera j. To simplify the notation, we assume that the feature vectors have the same dimensionality across cameras. The action classifier $\mathcal{W}_a$ for action a is a $(C * d)$-dimensional classifier. For each frame i, the latent camera importance variable A_i is also a $(C * d)$-dimensional vector, where $A_i = [A_i^1, A_i^2, ..., A_i^j, ..., A_i^C]$, $j \in \{1 : C\}$. For each camera j, $A_i^j = \alpha_i^j * \mathbf{1}^d$ is a d-dimensional indicator vector that ensures all the dimensions of the feature vector corresponding to a camera are weighted equally.

Our bi-linear max margin model searches for the best camera importance variables that maximize the action prediction accuracy by:

$$\min_{A,\mathcal{W},\xi} \sum_{a=1}^{m} \parallel \mathcal{W}_a \parallel_2^2 + \lambda \sum_{i=1}^{N} \xi_i^a \tag{1}$$

such that

$$(A_i^T \odot \mathcal{W}_{y_i})^T (A_i \odot \mathcal{X}_i) > (A_i^T \odot \mathcal{W}_a)^T (A_i \odot \mathcal{X}_i) + 1 - \xi_i^a \quad \forall a \neq y_i, i$$

$$A_i = [A_i^1, A_i^2, ..., A_i^C]$$

$$A_i^j = \alpha_i^j * \mathbf{1}^d \quad j \in \{1 : C\}$$

$$\sum_{j=1}^{C} \alpha_i^j = 1, \quad \alpha_i^j \in [0, 1], \quad \xi_i \geq 0 \quad \forall i,$$

where $\mathcal{W}$ is the matrix of all action classifiers across all cameras ($\mathcal{W} = [\mathcal{W}_1 \mathcal{W}_2 ... \mathcal{W}_m]$), and ξ is the standard slack variable in max margin formulations.

The first constraint encourages the model to make correct predictions. This constraint pushes for (W, A) of an action to score higher for instances of that action compared to any other action model. The other constraints push the latent variable to be similar for all dimensions within a camera, and contributions of cameras form a convex combination.

Bi-linear relations between the importance variables and action classifiers does not allow direct applications of standard latent max margin methods [36]. To optimize for $A, \mathcal{W}, \xi$ we use block coordinate descent. This involves estimating $\mathcal{W}$ for fixed A and optimizing for A given fixed $\mathcal{W}$. Optimizing for $\mathcal{W}$ given a fixed A reduces to a standard max margin model and can be solved with quadratic programming in dual. We initialize $\mathcal{W}$ by independently trained classifiers for each action. We calibrate these classifiers using the methods of [37].

To encode higher-order correlations in the feature space, we also consider different combinations of cameras, where each combination of cameras can be though of as a new dummy camera. Section 4 shows the benefits of considering such higher order correlations via camera combinations.

4 Experiments

We evaluate our model on how accurately it can predict actions in the settings where an egocentric camera plus multiple static cameras observe human activ-

Approach	Avg. Acc.	Avg. per Class Acc.
Latent Dynamic CRF [38]	41.55	33.57
Hidden Unit CRF [39]	24.13	26.22
Our Method	**54.62**	**45.95**

Table 1. The comparison of our model with two of the state-of-the-art latent CRF models: Latent Dynamic CRF and Hidden Unit CRF. Our model outperforms both of these methods.

ities. We compare our method with state-of-the-art methods for multi-view activity recognition, such as the Latent Dynamic CRF and the Hidden Unit CRF, and several baselines that aim at evaluating different components of our model. To qualitatively evaluate the quality of the learned camera indicator variables (α), we utilize them in a virtual cinematography task.

In this paper, we are interested in learning to predict human actions in the presence of one egocentric camera plus multiple static cameras. Our experiments are designed to support this task.

4.1 Multiple Static and an Egocentric Camera

Dataset: We chose the challenging CMU Multi-Modal Activity dataset (CMU-MMAC) [29] because it has multiple static and one egocentric videos of subjects performing kitchen activities. We use brownie-making videos, because frame-level annotations of actions are provided. 11 out of 39 actions of the dataset are "unique" to different subjects, therefore it is impossible to recognize those actions with leave-one-subject-out cross validation. After discarding videos of subjects that have synchronization problems due to dropped frames in some cameras (in order to use all 4 cameras) and removing unique actions, we obtain a dataset of 28 different actions, for 5 different subjects, recorded from one egocentric and 3 static cameras in 1/30 sample rate. The final actions include close fridge, crack egg, open brownie bag, pour brownie bag into big bowl, pour oil into measuring cup small, twist on/off cap, stir, take fork, walk to fridge, switch on, open drawer and more. We will make the list of dropped frames and the list of subjects for whom all four cameras can be used publicly available for the CMU-MMAC dataset.

Features: Similar to the creators of the CMU-MMAC dataset [30], we use the GIST [40] features for all the methods and baselines in our experiments on the CMU-MMAC dataset. Eight orientations per scale and four blocks per direction resulting in 512 dimensional GIST features are used with PCA (99 % of data coverage: 80-121 dimensional features). In another study Taralova et al. [4] used STIP features and showed improvements over GIST features when using only the egocentric data in a bag of words approach (average precision of 0.734 vs. 0.478). In their setting they merged similar action categories into one class and ended up with 15 categories (vs. 28 categories in our experiments) and used 14 subjects for training and 2 for testing having 4 random disjoint sets (vs.

our use of 4 subjects for training and a fifth for testing). We choose to use GIST features, because they were originally used on the CMU-MMAC dataset and are suitable for making the comparisons of methods in this paper, where emphasis is not on the feature engineering.

Experimental Setup:

Our action model uses a sampling, where negatives are sampled in a 1:3 ratio. The same sampling is preserved across all comparisons. We use leave-one-subject-out as our experimental protocol and both average accuracy and average per class accuracy (Avg. Acc. and Avg. per Class Acc.) as our evaluation metrics. Our model achieves an average accuracy of 54.62 and average per class accuracy of 45.95.

According to our knowledge, this paper is the first attempt to fuse the information from an egocentric and multiple static cameras. We are the first to use the data from all 4 cameras from the CMU-MMAC dataset. Therefore, there is no existing baseline using the egocentric and static views of CMU-MMAC. For this reason, we compared our model with the existing state-of the art methods such as different latent CRF's besides providing our own baselines by modifying different parts of the model.

Comparisons to State-of-the-Art Methods: We compared our method with state-of-the-art methods in multi-view, temporal classification. We selected state of the art methods that are applicable to the settings of an egocentric and multiple static cameras. Note that methods that rely on 3D estimates of the visual hull of the subjects are not directly applicable to egocentric cameras.

Different versions of latent model CRFs have been successfully used in multi-view action recognition. In particular, we compare our method with the Latent Dynamic CRF and the Hidden Unit CRF. The Latent Dynamic CRF [38] is a discriminative method with a strong track record for multi-view activity recognition. It models the sub-structure of action sequences by introducing hidden state variables. In our experiments the best results are obtained by using one hidden node per label. Table 1 shows that our model outperforms the Latent Dynamic CRF on the challenging task of action recognition with one egocentric and multiple static cameras. We also compare our model with the Hidden Unit CRF [39] where there are hidden nodes between action classes and features. Those hidden nodes can reveal the latent discriminative structure in the features. For each frame a Hidden Unit CRF can represent nonlinear dependencies. The best results in our experiments are obtained by using a total number of 100 hidden units. Table 1 shows that our method also outperforms the Hidden Unit CRF.

Both the Latent Dynamic CRF and the Hidden Unit CRF approach the problem of action recognition by joint reasoning over time. In the case of combining egocentric and static cameras, coupling joint temporal reasoning with discovering the latent structure in the high-dimensional feature space makes the problem extremely challenging. We postulate that separating temporal reasoning from discovering the latent structure might result in more accurate estimates of the latent structure.

Approach	Avg. Acc.	Avg. per Class Acc.
Baseline 1	41.80	44.05
Baseline 2	52.00	41.55
Baseline 3	47.93	42.45
Baseline 4	44.58	35.01
Baseline 5	48.83	45.89
Baseline 6	43.55	39.81
Baseline 7	48.75	45.66
Baseline 8	35.04	36.97
Our Method	**54.62**	**45.95**

Table 2. We compare our method with several baselines. Removing temporal smoothing, considering binary alphas, not considering camera combinations, and encoding per-action α hurts the performance of our model. This supports our intuitions about different parts of our model.

Baselines: To further analyze our model, we examine the effects of each component in our model with several baselines designed to challenge different components in our formulation. Except for baseline 1, which measures the effect of Viterbi smoothing, all other baselines use a final Viterbi smoothing stage.

Baseline 1: To examine the importance of encoding temporal information, we remove the Viterbi temporal smoothing at inference and compare it with our full model. Table 2 shows that encoding temporal information helps action recognition in our setting.

Baseline 2: To verify the effects of binary versus continuous α (camera selection vs using all cameras with respect to their calculated importance), we train our model with the binary α constraint, where $\alpha_i^j \in \{0, 1\}$. This forces our model to pick only one camera combination. Table 2 shows results when binary α is used.

Baseline 3: To challenge the observation about encoding higher-order correlations using camera combinations, we compare our method with a version that uses only four cameras (no combination). This baseline uses continuous α. Results in Table 2 show that leveraging higher-order relations of cameras (camera combination) in our latent discriminative model improves action recognition.

Baseline 4: This baseline is similar to the previous one, with one modification: using binary α (camera selection). Table 2 shows that both higher-order correlations of features (camera combination) and using all cameras according to their importance improves the recognition accuracy.

Baseline 5: Our optimization learns an α for each frame. One could reasonably worry about a large number of parameters to learn. An alternative is to search for an α for each action. This implies that the importance of the cameras are fixed for all frames during the course of an action. Baseline 5 corresponds to experiments with a per-action α model. The results in Table 2 support our intuition that the importance of the cameras changes during an action.

Camera Combinations	Avg. Acc.	Acc. Drop %
Ego + Side + Back + Top	44.58	0
No Side	42.97	3.6
No Top	40.56	9.0
No Back	43.54	2.3
No Ego	**27.99**	**37.2**

Table 3. To evaluate the importance of each camera in our formulation we remove a camera from our model (with binary α) one at a time, and report the drop in the accuracy. The egocentric camera is the most informative camera in this setting and the back static camera is the least useful one.

Baseline 6 examines the need for latent variables while learning to recognize actions across multiple cameras. In this baseline, we fuse multiple cameras without the latent variable. This late fusion baseline uses action models trained independently for different cameras and fuses them by a second layer RBF SVM. This baseline uses higher-order correlation of features(camera combination) and Viterbi smoothing. Our model differs from this baseline in using the latent α to explicitly encode the relationships across cameras in a discriminative manner. Table 2 shows the importance of our latent variable.

Baseline 7: Instead of learning latent variables, this late fusion baseline combines multiple cameras by equal weights and uses Viterbi smoothing.

Baseline 8: Another way to combine multiple cameras is to combine all the observations at the feature level and expect the classifiers to discover complex relationships in this high-dimensional data. Baseline 7 corresponds to this early fusion model. The results in Table 2 imply that complex relationships cannot be reliably discovered by just fitting a classifier to the combination of features. This baseline also uses Viterbi smoothing.

Besides the given state-of-the-art models and the eight baselines, we also experiment with per-frame classifiers of Logistic Regression (LR) and Nearest Neighbor (NN), and both underperform our model (Avg. Acc. LR=31.1, NN=37.1, c.f. 54.6 of ours).

As a sanity check, we also determine if there is apparent discriminative information in any single cameras that can bias the recognition performance on the CMU-MMAC dataset. To do that, we train RBF-SVMs action classifiers with Viterbi using only one camera. Using egocentric, static back, static side, and static top cameras alone results in average accuracies of $37.92, 22.07, 21.26, 20.86$, respectively (compared to average accuracy of 54.62 using our full model).

Although our main purpose is not to determine the best type of camera and position in recognizing actions, in order to further explore the importance of each camera in our model, we remove one camera at a time using binary α, and report the percentage drop in the performance numbers. Table 3 compares the accuracies and also the percentage drop in removing each of the cameras.

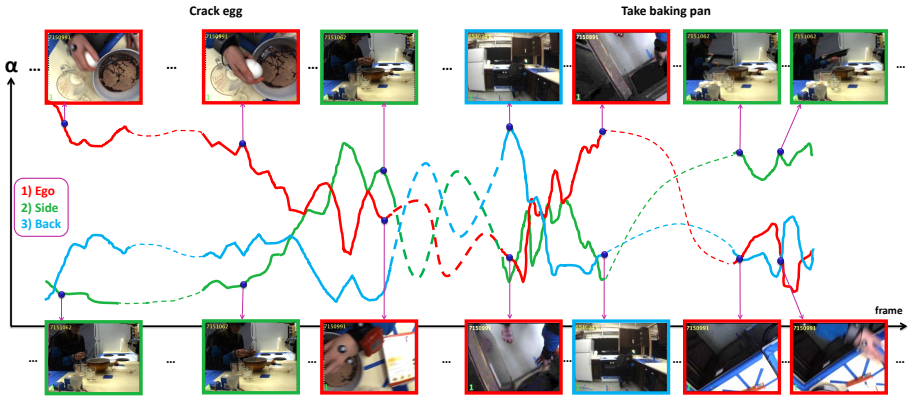

Fig. 2. The distribution of the learned α for "crack egg" and "take baking pan" actions: Our model learns that the egocentric camera contains maximum information for cracking an egg (highest α) when the subject interacts with the egg. Toward the end of the action when the subject looks away from the action, our model assigns more weight to the side static camera that represent the action best. For the "take baking pan" action our model allocated more weight to the static back camera when the subject walks to the cabinet. After opening the cabinet door, our model assigns more weights to the egocentric camera. This is followed by putting more weights on the side camera, when the subject is about to put the pan on the counter top.

The most informative camera in our setup is the egocentric one, and the least informative one is the back camera. This result is consistent with the nature of the dataset, where most of the actions that require hand-object interactions are best encoded in the subject's viewpoint using an egocentric camera.

To qualitatively evaluate the learned α's, we depict the distribution of α for some frames belonging to the "crack egg" and "take baking pan" actions in Figure 2. When the subject is cracking the egg, our method assigns high weights to the egocentric camera and once the subject's head is turning (and the egocentric camera is not informative) then our method assigns more weights to the side static camera. For the "take baking pan" action our method assigns more weights to the back camera when the subject is moving toward the cabinet. Once the cabinet door is open and the subject starts searching for the baking pan, the egocentric camera claims more weight. Toward the end of this action, the side static camera becomes more informative, and our method assigns more weight to that camera.

Figure 3 shows our model's preference for cameras for each action. Actions like closing and opening fridge, cracking an egg, pouring oil into bow or measuring cup are better encoded by the egocentric camera. For actions such as walking to the fridge, taking a brownie box, and putting a baking pan into the oven, our model prefers the side view camera where the body movements are visible.

Figure 4 shows the confusion matrix we obtained from our experiments on the CMU-MMAC dataset. The numbers are rounded and represent percentages.

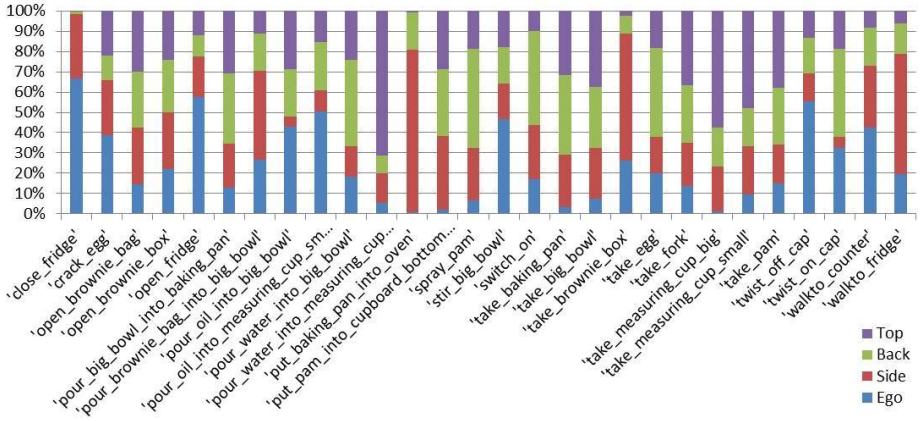

Fig. 3. The preference of our model in terms of cameras for each action. Actions like closing and opening fridge, cracking an egg, pouring oil into bow or measuring cup are better encoded in egocentric camera. For actions such as walking to the fridge, taking a brownie box, and putting a baking pan into oven our model prefers the side view camera where the body movements are visible.

Several off-diagonal confusions are due to the granularity of the action labels in the dataset. For example, the biggest confusions are between "take big bowl" and "take measuring cup small" or between "take measuring cup small" and "take measuring cup big". There are also confusions between actions that correspond to pouring either water or oil into a container.

Figure 5 interprets the confusions in a different way. It shows per-class action recognition accuracies and the most confusing action for all actions in the CMU-MMAC dataset. Like the confusion matrix in Figure 4 shows, the most confusion comes from very similar actions. The top three actions in terms of their recognition accuracy are "take egg", "put baking pan into oven", and "pour water into big cup". The three most difficult actions for our model are "twist off cap" and "put pan into cupboard bottom right", and "open brownie bag". Another source of confusion stems from the fact that some actions share very similar settings. For example, "open fridge" and "take egg" are frequently confused because the majority of the scene in the "take egg" action corresponds to the half-open door of the fridge. In addition, some actions share very similar body movements. For example, reaching for the same cabinet to take a PAM or a baking pan.

In addition to the experiments with the CMU-MMAC dataset, we have also run some experiments with static cameras on the IXMAS dataset, where the performances of the other methods are already available. Our method performed on-par-with the state of the art on the IXMAS dataset w/o any fine tuning. The details will be reported in a future paper.

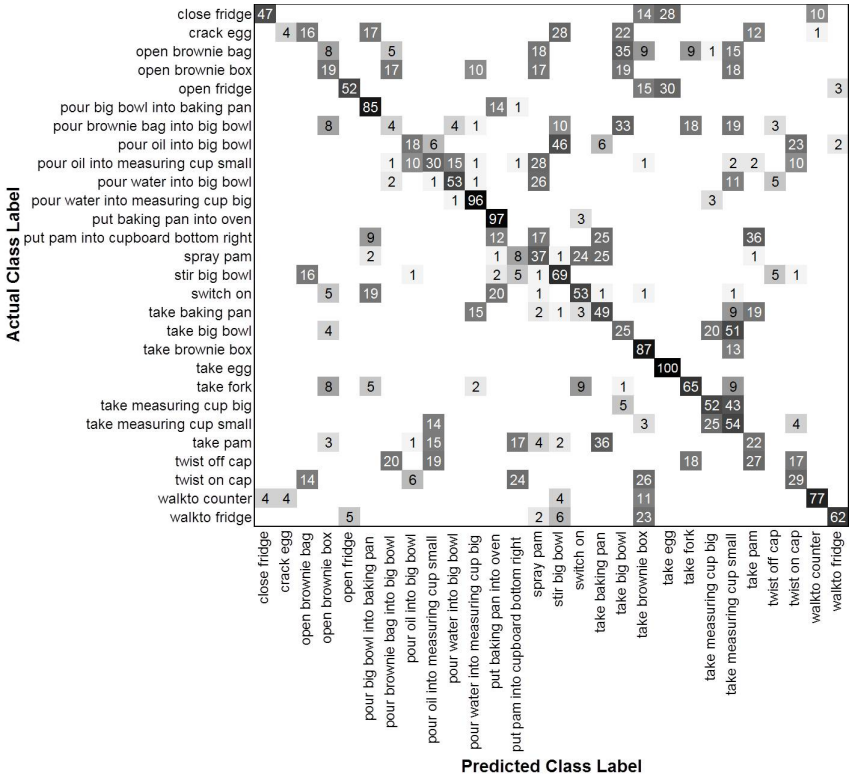

Fig. 4. Confusion matrix resulting from the experiments on the CMU-MMAC dataset. The numbers are rounded and represent percentages. Several off-diagonal confusions are due to the granularity of the action labels in the dataset. For example, the biggest confusions are between "take big bowl" and "take measuring cup small" or between "take measuring cup small" and "take measuring cup big". There are also confusions between actions that correspond to pouring either water or oil into a container.

4.2 Using α to Direct a Video Scene

To qualitatively evaluate our inferred α, we also use them to direct the scene of "making brownies" recorded from multiple static and an egocentric camera from CMU MMAC dataset. The task is to select which cameras to use for each frame. We use our learned α to select one camera per frame. Picking the camera with maximum α value results in undesirable frequent switches between cameras. To avoid that we use segmented least squares to smooth the camera transitions. Figure 6 shows examples of the frames across 3 cameras: static side, static back, and egocentric camera. Frames with red boxes are selected by our method. It is interesting to see that when the subject searches for the fork in a drawer, our method switches to the egocentric camera where all the items inside the drawer are visible (Image B in Figure 6). When the subject moves toward the cabinet

to take a baking pan, our method switches to the back camera where human movements are clearly visible (Image C in Figure 6). When the subject switches on the oven, our method picks the side static camera where the extended arm of the subject is visible (Image D in Figure 6).

The video in the supplementary material shows examples of the results of directing a scene using the learned α in our model. To avoid undesirable frequent camera switches, we use segmented least squares with 10-20 frames for smoothing. We do not switch between the cameras for inconsistent actions. We encourage the readers to watch the video in the supplementary material.

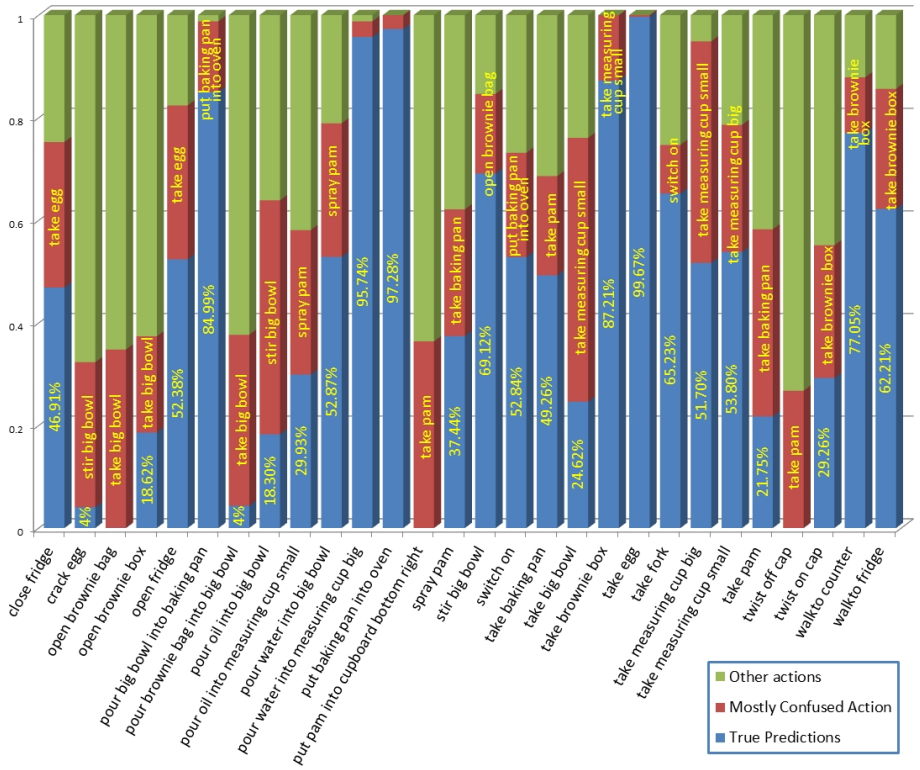

Fig. 5. Analysis of the action recognition on the CMU-MMAC dataset.

5 Conclusion

We introduce a model that, for each query frame, discriminatively predicts the importance of different cameras and fuses the information accordingly. We show that our model outperforms state-of-the-art methods that jointly reason across

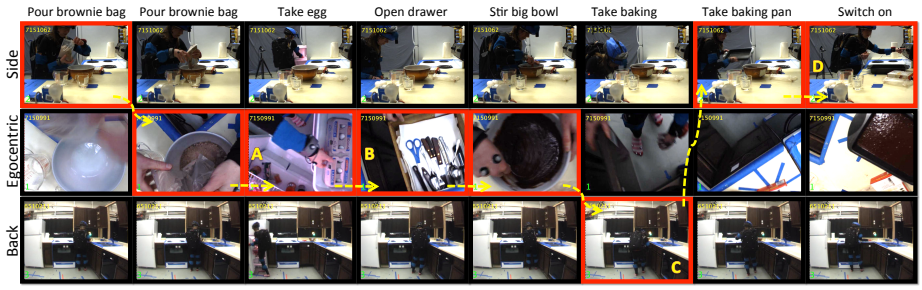

Fig. 6. This figure depicts some frames across 3 cameras (side, ego, back) for eight different actions. Red boxes indicated the frames that our method selects for the virtual cinematography. It is interesting to see that for the "take egg" action our model chooses the egocentric camera where one can clearly see the eggs (Image A). For "taking a fork", our model also switches to the egocentric camera where one can see the items inside the drawer (Image B). When the subject walks to the cabinet our model switches to the back camera (Image C). When the subject is about to turn the oven on our model picks the side camera, where the extended arm of the subject is visible (Image D). Please see supplementary material for the resultant videos.

time and actions. Our hypothesis is that joint reasoning across camera importance and actions followed by temporal smoothing is a more manageable learning problem than joint reasoning over time. By being more focused on frame-level discrimination, our model learns meaningful latent variables and can discriminatively suggest the importance of each camera for each frame. The next step involves extending our explicit latent variable formulation to also perform joint reasoning over time. Our learned camera indicator variable provides a level of understanding of the scene that enables meaningful camera selection for automatic cinematography. Our model does not take into account the principles of cinematography.

Our method shows very successful results on the challenging CMU-MMAC dataset by fusing the information from the egocentric and static cameras as needed. To the best of our knowledge, this is the first attempt in combining egocentric action recognition and conventional static camera action recognition.

References

1. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing (2010)
2. Ren, X., Philipose, M.: Egocentric recognition of handled objects: Benchmark and analysis. In: CVPR. (2009)
3. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. (2011)
4. Taralova, E., De la Torre, F., Hebert, M.: Source constrained clustering. In: ICCV. (2011)
5. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: ICCV. (2011)
6. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: ECCV. (2012)
7. Kanade, T., Hebert, M.: First-person vision. Proceedings of the IEEE (2012)
8. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR. (2012)
9. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: CVPR. (2012)
10. Kitani, K.M.: Ego-action analysis for first-person sports videos. IEEE Pervasive Computing (2012)
11. Ogaki, K., Kitani, K.M., Sugano, Y., Sato, Y.: Coupling eye-motion and ego-motion features for first-person activity recognition. In: CVPR Workshops. (2012)
12. Sundaram, S., Mayol-Cuevas, W.: What are we doing here? egocentric activity recognition on the move for contextual mapping. In: ICRA. (2012)
13. Sundaram, S., Mayol-Cuevas, W.: High level activity recognition using low resolution wearable vision. In: CVPR. (2009)
14. Li, Y.L., Fathi, A., Rehg, J.M.: Learning to predict gaze in egocentric video. In: ICCV. (2013)
15. Fathi, A., Rehg, J.M.: Modeling actions through state changes. In: CVPR. (2013)
16. Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? In: CVPR. (2013)
17. Farhadi, A., Tabrizi, M.K., Endres, I., Forsyth, D.A.: A latent model of discriminative aspect. In: ICCV. (2009)
18. Wu, X., Jia, Y.: View-invariant action recognition using latent kernelized structural svm. In: ECCV. (2012)
19. Gu, C., Ren, X.: Discriminative mixture-of-templates for viewpoint classification. In: ECCV. (2010)
20. Song, Y., Morency, L.P., Davis, R.: Multi-view latent variable discriminative models for action recognition. In: CVPR. (2012)
21. Wu, C., Khalili, A.H., Aghajan, H.: Multiview activity recognition in smart homes with spatio-temporal features. In: ACM/IEEE International Conference on Distributed Smart Cameras. (2010)
22. Weinland, D., Özuysal, M., Fua, P.: Making action recognition robust to occlusions and viewpoint changes. In: ECCV. (2010)
23. Farhadi, A., Tabrizi, M.K.: Learning to recognize activities from the wrong view point. In: ECCV. (2008)
24. Liu, J., Shah, M., Kuipers, B., Savarese, S.: Cross-view action recognition via view knowledge transfer. In: CVPR. (2011)

25. Huang, C.H., Yeh, Y.R., Wang, Y.C.F.: Recognizing actions across cameras by exploring the correlated subspace. In: ECCV. (2012)
26. Junejo, I., Dexter, E., Laptev, I., Perez, P.: View-independent action recognition from temporal self-similarities. PAMI (2011)
27. Gkalelis, N., Kim, H., Hilton, A., Nikolaidis, N., Pitas, I.: The i3dpost multi-view and 3d human action/interaction database. In: CVMP. (2009)
28. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. CVIU (2006)
29. De la Torre, F., Hodgins, J., Montano, J., Valcarcel, S., Macey, J.: Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. Technical report, CMU, RI (2009)
30. Spriggs, E.H., De la Torre, F., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: IEEE Workshop on Egocentric Vision. (2009)
31. Fisher, R., Reddy, P.: Supervised multi-modal action classification. In: Technical report, Carnegie Mellon University. (2011)
32. McCall, C., Reddy, K.K., Shah, M.: Macro-class selection for hierarchical k-nn classification of inertial sensor data. In: PECCS. (2012)
33. Zhao, L., Wang, X., Sukthankar, G., Sukthankar, R.: Motif discovery and feature selection for crf-based activity recognition. In: ICPR. (2010)
34. Elson, D.K., Riedl, M.O.: A lightweight intelligent virtual cinematography system for machinima production. In: AIIDE. (2007)
35. He, L.w., Cohen, M.F., Salesin, D.H.: The virtual cinematographer: a paradigm for automatic real-time camera control and directing. In: Computer graphics and interactive techniques. (1996)
36. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI (2010)
37. Scheirer, W.J., Rocha, A., Michaels, R., Boult, T.E.: Meta-recognition: The theory and practice of recognition score analysis. PAMIs (2011)
38. Morency, L., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: CVPR. (2007)
39. van der Maaten, L.J.P., Welling, M., Saul: Hidden-Unit Conditional Random Fields. In: IJCAI. (2011)
40. Oliva, A., Torralba, A.: Building the gist of a scene: the role of global image features in recognition. In: Progress in Brain Research. (2006)