Recognizing Activities with Multiple Cues

Rahul Biswas[1], Sebastian Thrun[1], and Kikuo Fujimura[2]

[1] Stanford University
{rahul,thrun}@cs.stanford.edu
[2] Honda Research Institute
kfujimura@hra.com

Abstract. In this paper, we introduce a first-order probabilistic model that combines multiple cues to classify human activities from video data accurately and robustly. Our system works in a realistic office setting with background clutter, natural illumination, different people, and partial occlusion. The model we present is compact, requires only fifteen sentences of first-order logic grouped as a Dynamic Markov Logic Network (DMLNs) to implement the probabilistic model and leverages existing state-of-the-art work in pose detection and object recognition.

1 Introduction

In recent years, there has been considerable success in bolstering the performance of object recognition by considering not just the object itself but also the context in which it occurs. For example, in [18] and [35], the recognition of objects is boosted by analyzing contextual information within a camera image, such as the presence and absence of other objects, the relative location of the object in question, and global features characterizing the scene. That work expresses the concept of context through probabilistic relationships of multiple recognizers. A probabilistic model posits the relationship between context and objects.

In this paper, we seek a plausible extension of this work to image sequences. By tying together information about peoples' poses, objects seen, and relative locations, we seek to identify peoples' activities using probabilistic models pertaining to these cues. Using all three cues in tandem offers far greater accuracy than any of the cues on its own.

To achieve these results, we develop a novel framework for expressing the spatio-temporal relation of cues and activities. Our framework is based on Dynamic Markov Logic Networks (DMLNs) [31], a well-established first-order probabilistic representation. We demonstrate in this paper that the DMLN framework provides a powerful language to express the probabilistic relationship of cues and activities in the video analysis domain. In fact, we show that useful DMLN theories establish the notion of context significantly more compactly than the propositional representations used, for example, in [18] and [35]. Our approach introduces new inference techniques for DMLN inference that accommodates the specific nature of the inference problem in the computer vision domain.

Our experiments apply state-of-the-art techniques for pose detection and object recognition. We show empirically that DMLNs effectively leverage context

A. Elgammal et al. (Eds.): Human Motion 2007, LNCS 4814, pp. 255–270, 2007.

and achieve improved recognition rates. These results are well in tune with earlier work on this topic. Hence, we conjecture that the DMLN framework provides an elegant and effective way to extend that work into the temporal domain involving human activities – a research topic that has found considerable attention in the computer vision field in past years [25], [16], [13], [7].

2 Overview of DMLNs

In this section, we present the concept of Dynamic Markov Logic Networks (DMLNs). While a formal treatment is far beyond the scope of this work, we refer the interested reader to [29] and [31], the definitive works regarding DMLNs.

A DMLN uses the language of First-Order Logic [10] to express a First-Order Probabilistic Model [23]. It explicitly posits the notion of objects and boolean predicates regarding them. Let us illustrate these notions by way of an example.

First, let p, q, and r be objects. These can be many things, depending on the problem of interest. Examples include an image, a single pixel, a feature being tracked, a robot, or a database entry. Probabilistic models such as Markov Random Fields or Dynamic Bayesian Networks consider only propositions, possibly about objects, but cannot consider objects explicitly.

Second, let $A(x, y, t)$ and $B(z, t)$ be fluents. Examples of fluents include whether one tracked object is occluding another at time step t or whether a specific pixel position in a video stream is part of an edge. The fluents relate to variables which will be substituted with actual objects at runtime.

Each possible assignment of objects to fluents forms what is known as a ground fluent. In this example, the ground fluents are $A(p, p)$, $A(p, q)$, $A(p, r)$, $A(q, p)$, $A(q, q)$, $A(q, r)$, $A(r, p)$, $A(r, q)$, $A(r, r)$, $B(p)$, $B(q)$, and $B(r)$. Each ground fluent is true or false. Let us consider a concrete example. Let p, q, and r be three people. Let $A(x, y, t)$ mean that x and y are friends at time t and let $B(z, t)$ mean that z is cheerful at time t. In this example, at time 0, let p and q be friends with one another and r is unknown to the other two. Moreover, let p be the only cheerful one. Then ground fluents $A(p, q, 0)$, $A(q, p, 0)$, $A(p, p, 0)$, $A(q, q, 0)$, $A(r, r, 0)$, and $B(p, 0)$ are all true and all other fluents at time 0 are false. We follow the arbitrary convention that one is always friends with oneself.

In addition to objects and fluents, DMLNs also have weighted sentences that give rise to a probability distribution over models, i.e. truth assignments to all ground fluents. Such a sentence might be

$$B(x, t) \rightarrow B(x, succ(t)) \tag{1}$$

with weight 2.0. It means that one typically but not always remains cheery in the future when one has been cheery in the past.

Another sentence might be

$$B(x, t) \wedge A(x, y, t) \rightarrow B(y, t) \tag{2}$$

with weight 1.0. It means that cheerful people pass on their cheerfulness to their friends. This effect is weaker than retaining one's own cheerfulness.

While DMLNs properly give rise to a joint probability distribution over ground fluents, let it suffice to state for this exposition that models that satisfy more sentences are more likely than those that do not. Moreover, models that satisfy higher weight sentences are more likely in a probabilistic sense than those that satisfy lower weight sentences.

3 Activity Recognition System

3.1 Probabilistic Model

The key contribution of this paper is to express a graphical model for combining multiple cues – pose, object presence, and movement location – through a compact DMLN theory. While graphical models have produced excellent results in computer vision applications, they have been too complicated – difficult to construct, modify, and communicate. We present our DMLN model in Figure 1 as an alternative.

1	$\forall t \forall a$ Activity(a,t) $\rightarrow$ Activity(a,succ(t))	8.3
2	$\forall t \forall a_1 \forall a_2$ Activity(a_1,t) $\rightarrow$ =(a_1,a_2) $\vee$ $\neg$ Activity(a_2,t)	∞
3	$\forall t \exists a$ Activity(a,t)	∞
4	$\forall t \forall a$ Activity(a,t) $\rightarrow$ Pose(a,t)	0.7
5	$\forall t$ Activity(a,t) $\wedge$ Useful(o,a) $\rightarrow$ Present(o,t)	0.7
6	Useful(TYPING,KEYBOARD)	∞
7	Useful(MOUSING,MOUSE)	∞
8	Useful(EATING,CANDY)	∞
9	Useful(EATING,APPLE)	∞
10	Useful(DRINKING,SODA)	∞
11	Useful(READING,BOOK)	∞
12	Useful(TALKING,PHONE)	∞
13	Useful(WRITING,PEN)	∞
14	Useful(WRITING,PAPER)	∞
15	$\forall t \forall a$ Activity(a,t) $\rightarrow$ Movement(a,t)	0.2

Fig. 1. Probabilistic Model as Dynamic Markov Logic Network (sentence weights appear in the right column)

3.2 Fluents and Sentences

Sentence 1 of Figure 1 refers to the predicates Activity(a,t) and Activity (a,succ(t)). Activity(a,t) is a binary variable that means that the person is engaged in activity a at time t. For example, Activity($WRITING$,1244) means that the observed person is writing in frame 1244 of the video sequence. We will assume throughout that there is exactly one person in each image. One could relax this assumption along the lines of [12] if need be.

This sentence states that activities tend to persist over time. In other words, if you are drinking from a can of soda, you will most likely continue to be still drinking from that can in the next frame, tens of milliseconds later. In the sentence, succ(t) means that successor to t, i.e. the next frame. The sentence literally reads – for all time steps t and for all activities a, if the user is engaged in activity a at time step t, then the user will be engaged in activity a in the time step following t.

Sentence 2 establishes a mutual exclusion constraint between simultaneous activities. Simply put, we are providing the machine a generalization of the notion that a person cannot walk and chew gum at the same time. This sentence reads – for all time steps t and for all activities a_1 and for all activities a_2, if the user is engaged in activity a_1 at time t, then either a_1 and a_2 refer to the same activity or the user is not engaged in activity a_2 at time t. This sentence has the special property that it has infinite weight, i.e. this is a hard constraint.

Sentence 3 states that the user must be doing something. It is possible to add a special activity, $NOTHING$, to indicate times where the user is not engaged in any activity.

3.3 Incorporating Cues

The output of the pose recognition system described in section 4 is represented by asserting the sentence Pose(a,t), which means that the person's observed pose appears to be in keeping with being engaged in activity a at time t. The pose detection posits a sentence weight reflecting its confidence, as do the other components. Sentence 4 states that a person's pose will reflect the activity they are currently engaged in.

The objects a person is using also provide valuables cues as to what they are doing. In this component, we are content to consider what objects are in view. If we see both a can of soda and a phone, we cannot definitely say which one is being used but if we do not see a keyboard, we are less likely to think that the person is typing.

The object recognition component uses the fluents Present(o,t) (an object of type o is present at time t) and Useful(a,o) (objects of type o are useful for activity a). Sentence 5 states that objects useful to an activity will be present when that activity is engaged in. Sentences 6 through 14 specify which objects are useful for which activities. Here, mousing refers to using a computer mouse and talking refers to conversing on a telephone.

Sentence 15 states that the activity undertaken influences the location of the movement in the camera image. The fluent $Movement(a,t)$ states which activity seems likely by analyzing movement alone.

Before we move on, let us ensure that we understand the sentences by understanding the effect of removing any one sentence in isolation. Removing the first sentence would eliminate the continuity of our temporal model, forcing the inference procedure to consider each image in isolation without the benefit of the entire video sequence. Removing the second and/or third sentence would allow for the classifier to choose no activity or to choose more than one. Removing the fourth sentence would prevent the pose detector from providing useful information and removing the fifth or fourteenth sentence would do the same for the object and movement detection, respectively. Removing any of the sentences from six through fourteen would make the system blind to the corresponding object type.

3.4 Ease of Expression in DMLNs

This paper is as much about building an activity recognition system using multiple cues as it is about the efficacy of DMLNs in computer vision. We use multiple cues and do activity recognition in this manner because it is robust against background clutter, lighting differences, intra- and inter-personal variance, and difficulties in pose and object recognition.

Our motivation for leveraging DMLNs is very different. We use the DMLN because it is possible to posit a temporal probabilistic model that is simple to express and easy to understand without compromising the sophistication necessary for high performance. Table 1 is our entire probabilistic model. We wrote a set of sentences that we thought represented activities and cues well and revised it until we were pleased with it. Contrast this with comparable models expressed with Markov Random Fields (MRFs) and Dynamic Bayesian Networks (DBNs). Those models are equally effective but they are hard to understand and challenging to implement.

In practice, DMLNs are inferentially equivalent to MRFs and DBNs. Indeed, to perform inference on the DMLN, we convert it twice, first to a ground MLN which is an MRF and then to a DBN. This process is straightforward except for the challenge that is addressed in section 5.1. Thus, one can think of DMLNs much as one would view a high-level computer programming language like C as opposed to a low-level assembly language. The study and design of systems in assembly language paved the way for the new languages. While initially the new language only brought programmer convenience, it eventually allowed for greater abstraction and better programs.

4 Obtaining the Ground Predicates

In section 3, we explored a theory for understanding how various activities give rise to cues that are useful in identifying those activities. In this section, we discuss how we use computer vision algorithms to translate our video stream into assertions regarding the cue predicates.

4.1 Pose Detection

Comparing Poses. Traditionally, pose detection is the problem of recovering the position of important body parts from a single image. For example, for a person playing baseball, a pose detection algorithm may identify the x,y positions in the image of her feet, hands, and head [17]. Or, for a person walking down the street, an algorithm may fit a skeleton, thus identifying the feet, knees, and hip [34].

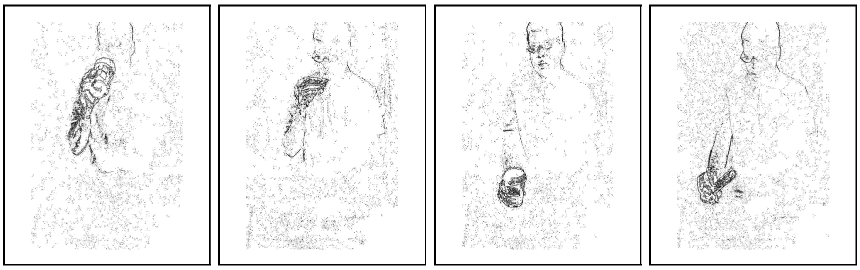

Fig. 2. Examples of difference images

We take an alternate approach to pose detection. The process of inferring body part positions and then linking configurations to activities is problematic on two levels. First, traditional pose detection is a difficult problem and even the best algorithms are subject to both inaccuracy and catastrophic failure. Second, the simplified 2D representation of pose is a poor indicator of human activity. Studying the evolution of pose over time is even less effective as the errors in pose detection create more phantom movements than actual ones. Earlier versions of our pose detection system suffered from each of the aforementioned difficulties.

We eliminate the intermediate step of recovering body part positions and instead map changes in pose to activities directly. To identify the change in pose, we start with difference image generated by subtracting two consecutive frames and thresholding it:

$$d_{x,y} = abs(i^{t}_{x,y} - i^{t+1}_{x,y}) \tag{3}$$

Examples of such difference images appear in Figure 2. Note that this difference image captures the outline of the body part that is moving, even if it does so imperfectly.

Second, to allow comparison between difference images, we construct a list of all pairwise point to point distances of the illuminated pixels in the difference image. This list retains the basic shape of the image while removing exact position and orientation information.

Third, we form a histogram over the list of distances just constructed. This is similar to how shape contexts [4] compact distance information except that angles are considered there but not here.

To compute the similarity of two histograms, we use the same scoring function as shape contexts:

$$\sum_{i=1}^{N} \frac{(h^1_i - h^2_i)^2}{h^1_i + h^2_i} \tag{4}$$

where N is the number of bins in the histogram.

Realizing the Pose Cue Predicates. We have established a metric for determining if two frames from different video streams represent the same or different changes in pose. To make this useful, we capture a several minute training video with different people performing each of the activities we are working to recognize. The new activity being undertaken is marked at activity transitions (e.g. when a person stops typing and starts drinking) in this video are labeled by hand. When the machine is later observing novel video, it compares the histogram of the incoming frame t^* with each of these histograms of the frames in the training video, as in [32]. All histograms are cached to ensure that this is a fast operation. For each activity, the number of comparisons exceeding a threshold is counted. The activity with the greatest count is selected (let it be a^*) and if the count exceeds a predetermined threshold, then the ground predicate $Pose(a^*, t^*)$ is observed to be true.

4.2 Object Detection

Object detection offers a far more out of the box output for activity recognition than does pose detection. To compare two images, we find all of the SIFT features from each of the images. Then we compute the number of features the two images share in common.

To realize object-related cues, we first crop by hand images of the objects used form the training video. Second, we label each image with the object inside that image (e.g. a pen).

When identifying objects in novel frames, the program compares the large novel frame with each of the cropped training images. For each object category o^* for which the novel frame t^* has over a fixed threshold of matches, the predicate $Present(o^*, t^*)$ is asserted. Note that the matches may come from different cropped images. For example, if the threshold is fifth SIFT matches and there are three cropped images of a candy bar with twenty-four, fifteen, and twelve SIFT matches respectively, then the program believes that a candy bar is present even though that could not have been ascertained from a single image.

4.3 Movement

We found that for some activities, movement tends to occur in some areas more than others. This pattern was previously noted in [35]. For example, mouse movement tends to occur in the bottom half of the screen. While this is a weak source of information on its own, it provides a valuable third stream of information to the probabilistic model. It can differentiate between drinking and using the mouse, for example, but not between drinking and eating. On its own, it fares rather poorly because the relative movement depends on camera placement relative to the subject and that was not rigidly fixed in our experiments.

We learn a mixture of Gaussians from training data about typical movement locations as expressed in the training data. Each activity is thus represented by the mean and covariance of observed movement locations of that activity in the training data. The posterior weights of novel frames are asserted as Movement(a,t).

5 Inference

As DMLNs are inferentially equivalent to DBNs [29], we can use efficient approximate inference algorithms designed for the latter. We use a variant of the standard Rao-Blackwellized Particle Filter (RBPF) [9] that runs the particles both forward and backward, as in [26].

A difficulty arises however. Sentences 2, 3, and 6 through 14 all have infinite weight and while this does not break inference, it does slow it down considerably. In our proposed DMLN, nearly all of proposed next states will be inconsistent and receive weights of 0.

What we have here is a mixed network – a probabilistic model with both probabilistic and deterministic components. Inference in mixed networks is well understood in atemporal settings [2], [3], [8], [15]. In handling DBNs generated from DMLNs though, we believe that generating a more compact, inferentially equivalent non-mixed DBN works best. For DMLNs to have broader applicability in computer vision past our specific application, we need a general algorithm to compact mixed DMLNs. We include such an algorithm here.

5.1 Compacting Mixed DMLNs

The process of converting a DMLN into a DBN is slightly tortuous and involves generating a Markov Random Field (MRF) as an intermediate step. It is at this stage we will compact the model. Let L be a single time slice of the MRF with $\bar{L}$ referring to deterministic potentials and $\hat{L}$ referring to probabilistic ones. Moreover, let $\bar{P}$ refer to all nodes referred to by deterministic potentials and let $\hat{P}$ refer to all nodes not referred to by deterministic potentials. Thus, if a node is referred to in both $\bar{L}$ and $\hat{L}$, it will be in the set $\bar{P}$ but not in the set $\hat{P}$.

First, we divide the nodes in $\bar{P}$ into independent subproblems, i.e. into maximally disjoint sets such that there exists no potential in $\bar{P}$ that refers to nodes in different subproblems. This is done by postulating a graph with MRF nodes as graph nodes and adding edges if and only if there exists a potential in $\bar{P}$ that refers to both nodes. The connected components of this graph are the independent subproblems of our MRF.

Second, for each independent subproblem, we use WalkSAT, a fast Constraint Satisfaction Problem sampling technique [39], to identify all solutions to the subproblem. This can take up to $O(2^N)$ time where N is the number of nodes that are linked together by deterministic potentials. This is clearly better than the non-compacted network, which will require at least that much time at every single time step whereas compacting requires it just once. In practice, WalkSAT takes only linear time in the number of solutions, barring pathological DMLNs.

Third, we replace each independent subproblem with a single multinomial variable, in the same manner as the variable elimination algorithm [24].

While the entire process of converting DMLNs into a form suitable for RBPF inference does have such intricacies, it is important to note that this pipeline is independent of the specific DMLN and application. Indeed, toolkits for DMLN inference are already publicly available [1].

5.2 Weight Learning

The algorithm in [31] learns weights directly from training data, freeing us from such considerations as how long people actually drink from a can of soda. It essentially follows a frequentist strategy, assigning a weight that corresponds to the observed probability of that sentence in the training data. The sentences appearing in the right-hand column of Figure 1 are from this algorithm. Higher weights indicate greater certainty with infinite weights indicating deterministic sentences.

6 Experimental Results

To evaluate our algorithm, we solicited volunteers for what they believed was a psychology experiment. When each volunteer arrived, they were seated at a desk in our laboratory and provided with a list of the following activities in a random order and with repetition. First, they were to write a paragraph with a pen and notebook. These and all other objects were placed on the desk at which the subjects were seated. Second, they skimmed several pages of a textbook. Third, they ate part of a candy bar and drank from a can of soda. Fourth, they answered a phone call. Fifth, they typed on a laptop and used an external mouse.

The subjects were seated at a wooden desk illuminated by sunlight. A Canon HV10 consumer grade video camera recorded the scene from atop a fixed tripod looking down on the scene. This viewpoint is common on many laptop cameras with built-in webcams as well as several off the shelf webcam mounting kits. Objects often remain in view when not being used and other objects clutter the desk where activities take place. The camera was set to focus and adjust light balance automatically. Audio input was not used.

Training data for the experiments was comprised of the same activities performed in the same manner. Each transition from one activity to another was marked by hand and this was used to generate activity labels for each frame. Also, four images of each object were cropped by hand. The training data includes twenty minutes of video. All frames had to be labeled as one of the seven activities.

The algorithm received no additional information about the test sequence except for the raw video stream. The metrics that follow are calculated on a frame by frame basis. The algorithm was forced to label all frames. The complete test sequence was eight minutes long.

6.1 Information Gain

Figure 3 presents the information gain of each component, that is, the reduction in entropy of the confusion matrices as different components are added to the DMLN. Here, an asterisk denotes that Sentence 1 from Table 1 was included in the DMLN. For reference, ground truth represents an information gain of 2.8.

Components	Information Gain
POSE	0.4
PRESENT	1.3
MOVEMENT	0.5
POSE *	1.7
PRESENT *	2.0
MOVEMENT *	0
POSE + PRESENT *	2.5
POSE + MOVEMENT *	1.8
PRESENT + MOVEMENT *	2.3
POSE + PRESENT + MOVEMENT *	2.6
Ground Truth	2.8

Fig. 3. Information Gain from Different Components

6.2 Confusion Matrices

Figure 4 shows a confusion matrix for pose detection. Confusing drinking and eating is the most common mistake and for good reason – we lift food to our mouths without regard to its phase. In mousing and talking, we hold roughly similar sized objects in our hands, casting a unique signature to the pose detection system. The difference in location only comes in when movement location is considered.

Figure 5 shows a confusion matrix for movement location detection. Movement fares poorly but the confusion matrix holds interesting clues. It works well for mousing, which is predominantly in the lower right of the image as all of our subjects were right handed. Activities such as writing, reading, and typing were found to be predominantly in the bottom half of the screen while drinking and talking were in the top half.

	WR	RE	EA	DR	TA	TY	MO
Writing	55	5	7	2	14	7	10
Reading	7	64	9	0	15	2	1
Eating	5	14	46	7	17	8	3
Drinking	15	12	32	9	13	5	13
Talking	21	8	3	2	40	17	9
Typing	11	2	24	3	27	26	8
Mousing	23	3	9	4	31	17	13

Fig. 4. Confusion Matrix for Pose Detection

6.3 Accuracy

Since many of the computer vision components are computationally expensive, the processing time is linearly dependent on both the resolution and frame rate. While lowering the resolution can make object recognition impossible, the frame

	WR	RE	EA	DR	TA	TY	MO
Writing	3	0	0	7	1	89	0
Reading	0	0	0	26	7	67	0
Eating	0	0	0	33	16	44	7
Drinking	0	0	0	81	1	18	0
Talking	0	0	0	69	22	9	0
Typing	1	0	0	0	0	99	0
Mousing	0	0	0	11	0	24	65

Fig. 5. Confusion Matrix for Movement Location Detection

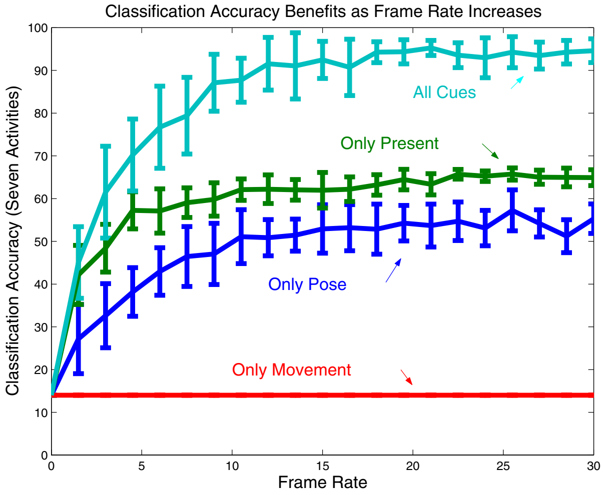

Fig. 6. Classification Accuracy as Frame Rate Increases

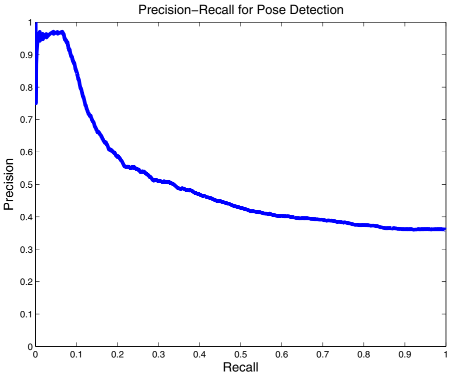

Fig. 7. Precision-Recall for Pose Detection

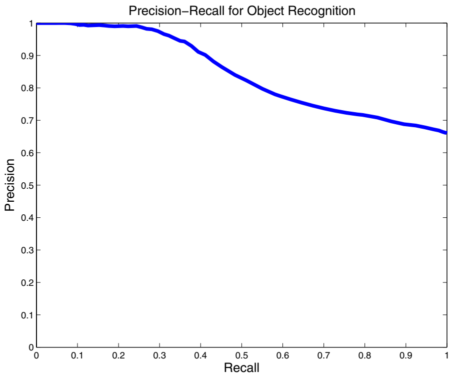

Fig. 8. Precision-Recall for Object Recognition

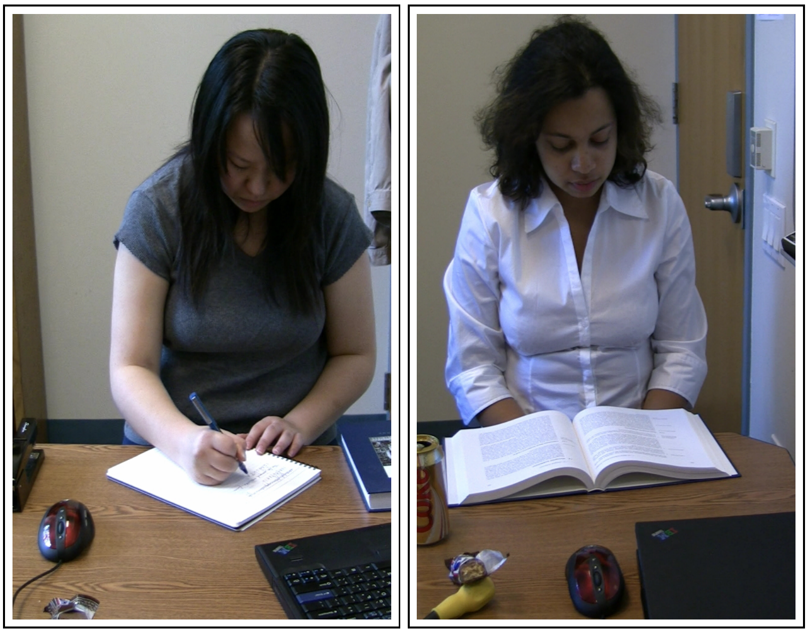

Fig. 9. An example where the pose detector correctly classifies the image but the object recognizer finds neither the pen nor the paper (left) and an example where the object detector finds the book but the pose detector misclassifies the activity (right)

rate provides a more balanced trade-off between computation time and performance. The effects of this balance is explored in Figure 6.

This figure also shows an alternate view of the value of each component after being processed by the DMLN as well as the benefit of bringing all the cues together. By itself, pose recognition could only classify 55% of the frames correctly; object recognition could only classify 65% of the frames correctly; and movement location could classify only 14% of the frames correctly. In tandem however, they are able to classify 95% of the frames correctly. This is the benefit of using multiple cues.

Figures 7 and 8 show the precision-recall graphs for pose detection and object recognition, respectively. These graphs highlight how far the field has progressed and the potency of [32] and [38]. They also show the benefit of using a DMLN as the latter incorporates information in proportion to the pose detection and object recognition algorithms' confidences.

Figure 9 show examples of where the pose detector works but the object recognizer fails and vice versa. This speaks to the need for multiple cues as no single cue will suffice in all situations.

7 Related Work

Pose detection seeks to identify the positions of a person's body parts from images without the use of motion capture devices or other artificial markers. It is a difficult problem in general because of differences in people's appearances, self-occlusion, self-shadows, and non-Lambertian clothing. Despite these challenges, several promising approaches have emerged [27], [22], [33], [28], [34], [32]. Three basic approaches typify the field. The first (e.g. [28]) takes a bottom-up approach, first identifying body parts and building them up into complete poses. The second (e.g. [34]) takes a top-down approach, performing joint optimization on the entire image at once. The third, introduced by [32], opts for a memory-based approach, recognizing pose by comparing images to existing training examples.

Object recognition and localization are equally difficult problems but we see substantial progress here as well. The challenges here include occlusion, illumination inconsistency, differences in viewpoint, intra-class variance, and clutter. Most techniques focus on either features [37], [38], [19], [30] or shapes [6], [21]. The advent of the SIFT descriptor [14] marked a great step forward for same-object recognition and most feature based approaches use SIFT or similar descriptors in a bag of words (e.g. [38]) or constellation (e.g. [11]) setting. Shape based approaches work well for different objects of the same category. The geometric blur feature descriptor [5] has proven especially effective here.

The use of multiple cues for object detection has primarily been explored by Torralba and colleagues in [20], [18], [35], [36]. They explore different probabilistic models in these works but the central theme throughout is to leverage hypothesized location information to place a prior over possible objects. They do this both to increase classification accuracy as well as improve running time. [16] follows a similar approach but focuses on integrating speech and pose.

Activity recognition [25], [16], [13], [7] has seen growing interest. [25] is similar to this work in their notion of activities but focuses on more complicated tasks (e.g. infant care) and uses RFID-tagged objects instead of cameras. [7] offers an elegant algorithm for detecting irregularities without requiring any labeling. [13] is also similar to this work in their notion of activities but focuses on substantial movement (e.g. going to the supermarket) and uses GPS data. None of these approaches however combine multiple cues, relying instead on a single source of information.

8 Future Work

Our results are promising and our framework makes it straightforward to extend our system with additional capabilities. This includes additional cues such as object location, perceived sound, and scene classification. We also want to evaluate our system on a non-stationary camera and in a wider variety of environments. Lastly, we want to explore integrating our system with a mobile robot.

9 Conclusion

In this paper, we introduced a first-order probabilistic model that combines multiple cues to classify human activities from video data accurately and robustly. The model we presented is compact, requiring only fifteen sentences of first-order logic grouped as a Dynamic Markov Logic Network (DMLNs) to implement the probabilistic model and leveraging existing state-of-the-art work in pose detection and object recognition.

Our results show that the algorithm performs well in a realistic office setting with background clutter, natural illumination, different people, and partial occlusion. It is robust against intra- and inter-person variance. We have shown promising results on classification accuracy, information gain, precision-recall, and confusion matrices.

References

1. http://alchemy.cs.washington.edu/
2. Allen, D., Darwiche, A.: New advances in inference by recursive conditioning. In: UAI 2003 (2003)
3. Bacchus, F., Dalmao, S., Pitassi, T.: Value elimination: Bayesian inference via backtracking search. In: UAI 2003 (2003)
4. Belongie, S., Malik, J., Puzicha, J.: Shape context: A new descriptor for shape matching and object recognition. In: NIPS 2000 (2000)
5. Berg, A.: Shape Matching and Object Recognition. PhD thesis, University of California, Berkeley, (Adviser-Jitendra Malik) (2005)
6. Berg, A., Berg, T., Malik, J.: Shape matching and object recognition using low distortion correspondence. In: CVPR 2005 (2005)
7. Boiman, O., Irani, M.: Detecting irregularities in images and in video. In: ICCV 2005 (2005)
8. Dechter, R., Mateescu, R.: Mixtures of deterministic-probabilistic networks and their and/or search space. In: UAI 2004 (2004)
9. Doucet, A., Freitas, N., Murphy, K., Russell, S.: Rao-blackwellised particle filtering for dynamic bayesian networks. In: UAI 2000 (2000)

10. Enderton, H.: A Mathematical Introduction to Logic. Academic Press, Inc, Florida (1972)
11. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. In: PAMI 2006 (2006)
12. Huang, C., Ai, H., Li, Y., Lao, S.: Vector boosting for rotation invariant multi-view face detection. In: ICCV 2005 (2005)
13. Liao, L., Fox, D., Kautz, H.: Extracting places and activities from gps traces using hierarchical conditional random fields. In: IJRR 2007 (2007)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2) (2004)
15. McAllester, D., Collins, M., Pereira, F.: Case-factor diagrams for structured probabilistic modeling. In: UAI 2004 (2004)
16. Morency, L., Sidner, C., Lee, C., Darrell, T.: The role of context in head gesture recognition. In: AAAI 2006 (2006)
17. Mori, G., Ren, X., Efros, A., Malik, J.: Recovering human body configurations: combining segmentation and recognition. In: CVPR 2004 (2004)
18. Murphy, K., Torralba, A., Freeman, W.: Using the forest to see the trees: A graphical model relating features, objects, and scenes. In: NIPS 2003 (2003)
19. Mutch, J., Lowe, D.: Multiclass object recognition with sparse, localized features. In: CVPR 2006 (2006)
20. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Progress in Brain Research: Visual Perception 155 (2006)
21. Opelt, A., Pinz, A., Zisserman, A.: A boundary fragment model for object detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, Springer, Heidelberg (2006)
22. Ormoneit, D., Black, M., Hastie, T., Kjellstrom, H.: Representing cyclic human motion using function analysis. In: IVC 2005 (2005)
23. Pasula, H., Russell, S.: Approximate inference for first-order probabilistic languages. In: IJCAI 2001 (2001)
24. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
25. Philipose, M., Fishkin, K., Perkowitz, M., Patterson, D., Fox, D., Kautz, H., Haehnel, D.: Inferring activities from interactions with objects. In: IEEE-PC 2004 (2004)
26. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition, pp. 267–296 (1990)
27. Ramanan, D., Forsyth, D., Zisserman, A.: Strike a pose: Tracking people by finding stylized poses. In: CVPR 2005 (2005)
28. Ren, X., Berg, A., Malik, J.: Recovering human body configurations using pairwise constraints between parts. In: ICCV 2005 (2005)
29. Richardson, M., Domingos, P.: Markov logic networks. Mach. Learn. 62(1-2) (2006)
30. Russell, B.C., Efros, A.A., Sivic, J., Freeman, W., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR 2006 (2006)
31. Sanghai, S., Domingos, P., Weld, D.: Learning models of relational stochastic processes. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, Springer, Heidelberg (2005)
32. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: ICCV 2003 (2003)

33. Sidenbladh, H., Black, M.: Learning the statistics of people in images and video. IJCV 54(1-3) (2003)
34. Sigal, L., Black, M.: Predicting 3d people from 2d pictures. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2006. LNCS, vol. 4069, Springer, Heidelberg (2006)
35. Torralba. A.: Contextual priming for object detection. International Journal of Computer Vision 53(2) (2003)
36. Torralba, A., Murphy, K.: Context-based vision system for place and object recognition. In: ICCV 2003 (2003)
37. Viola, P., Jones, M.: Robust real time object detection. In: SCTV 2001 (2001)
38. Wang, S., Quattoni, A., Morency, L., Demirdjian, D., Darrell, T.: Hidden conditional random fields for gesture recognition. In: CVPR 2006 (2006)
39. Wei, W., Erenrich, J., Selman, B.: Towards efficient sampling: Exploiting random walk strategies. In: AAAI 2004 (2004)