The Multi-level Learning and Classification of Multi-class Parts-Based Representations of U.S. Marine Postures

Deborah Goshorn, Juan Wachs, and Mathias Kölsch*

MOVES Institute
Naval Postgraduate School
Monterey, CA, USA

Abstract. This paper primarily investigates the possibility of using multi-level learning of sparse parts-based representations of US Marine postures in an outside and often crowded environment for training exercises. To do so, the paper discusses two approaches to learning parts-based representations for each posture needed. The first approach uses a two-level learning method which consists of simple clustering of interest patches extracted from a set of training images for each posture, in addition to learning the nonparametric spatial frequency distribution of the clusters that represents one posture type. The second approach uses a two-level learning method which involves convolving interest patches with filters and in addition performing joint boosting on the spatial locations of the first level of learned parts in order to create a global set of parts that the various postures share in representation. Experimental results on video from actual US Marine training exercises are included.

1 Introduction

The ability to automate the evaluation of human performance in training exercises using computer vision and behavior analysis is of recent interest in several research fields. It is a complex goal, but a building block of this goal is to create computer vision algorithms to detect the atomic events seen in training exercises. This paper is a result of yielding the fundamental posture recognition computer vision algorithms to support the automation of evaluating US Marines in their training exercises.

The four fundamental postures of a US Marine in training are the four torso orientations as illustrated in Figure 1. Significant posture changes are detected in order to evaluate high-level behavior anlaysis of Marines in training[9]. To recognize the four *object types*, or postures, this paper investigates the parts-based object representations and the multiple levels of learning of the parts that represent each posture in order to obtain robust representations of postures.

* The authors greatly appreciate and acknowledge the invaluable contribution by Noah Lloyd-Edelman with the video dataset.

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 505–512, 2009.

Fig. 1. Example training image from each object (US Marine posture) type

1.1 Related Work

This work is the first attempt to model multi-class human postures using parts-based approach on uniformed soldiers in a cluttered environment. It is a challenging problem, and this paper investigates the possibility of using multi-level learning of parts-based representation for multiclass body postures of uniformed soldiers to overcome the challenges of representing such diverse possible appearances of one posture class. This work is not the first attempt to represent objects as parts-based representations. In [1], parts representation of faces are learned using Multiple Cause Vector Quantization and Multiple Cause Factor Analysis, which are similar to Principle Component Analysis and Non-negative Matrix Factorization. In [3] and [4], a Bayesian approach to learning parts-based representations of objects is presented. The most related to this presented work is that of [2] and [5]. The second classifier presented is a variant of [2]. The first classifier presented in this paper is a multi-class version similar in learning, but different in classification to the pedestrian detector in [5].

2 Multi-level Learning and Classification of Parts-Based Object Representations

In order to use represent postures, or more generally, *object types*, with sparse representations, there must first be an attempt to learn the parts that make up the representation. It is advantageous to learn parts progressively, requiring multiple levels of learning. At each higher level of learning, the learning algorithm is more sophisticated, and the part learned is more sophisticated in its representation of the object type. For example, at the lowest level of learning, unsupervised learning like clustering learns salient features, or parts, from unlabeled parts data, that sparsely represents the object in common to the data. However, learning parts at this level is not always enough. The more levels of learning, the more sophisticated the parts-based representations are of the data.

The levels of learning may involve either purely one type of object or all types of objects. In the former, parts-based representations are learned for the sole purpose of object detection of that object type. There is no learning of parts that are discriminant between other object types. The second type of higher levels of learning which involve all object types, learn parts of object types that discriminate between object types. In this case, object types may even share parts in their individual parts-based representations. In this section, example parts-based recognition classifiers from both approaches are presented.

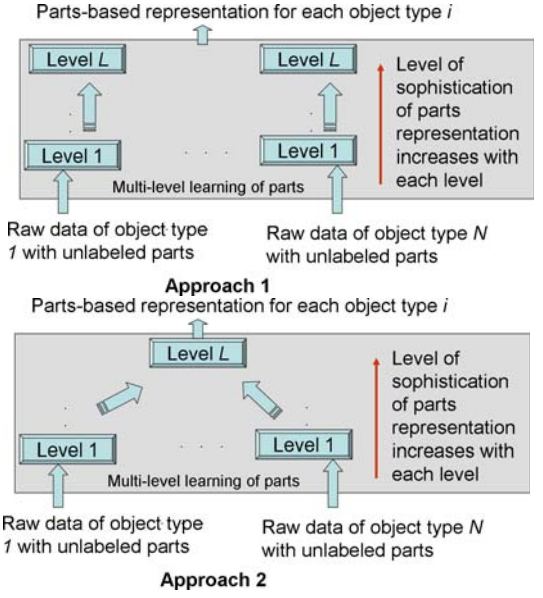

Fig. 2. Two approaches to multi-level learning of parts-based object representations

2.1 Approach 1 Example Parts-Based Object Recognition Classifier

One classifier presented in this paper is an example of a parts-based object recognition classifier that learns its parts in two levels, where all levels learn using data from one object type only. In other words, it is an instance of *Approach 1* with $L = 2$ from Figure 2. This classifier is similar to [5] in the learning part, but different in the classification process as soon described.

Level One Learning of Parts. Level one learning is entirely unsupervised with respect to learning the *parts* that represent an object type. For this classifier approach, the level one learning module as depicted in 3, inputs N_1 gray-scale training images of one object type 1, and outputs a *dictionary*, or set, of parts representing that particular object type. This learning level involves three majors steps. First, the Harris corner detector [8] is used to find the interest points on the object type attempted to be learned. Secondly, the interest patch, or window of size p x p extracted around each interest point, is collected from each training image. Here, p is fixed to 9, a mid-range value of standard patch sizes. Finally, a clustering algorithm, like the $K-$means clustering algorithm, is used to cluster the extracted patches (from all training images of a posture type) into K_1 clusters, where K_1 is the number of parts selected to comprise the dictionary for object type 1. The dictionary of *parts* is simply the resulting cluster means. In other words, if object part i is denoted as P_i, a p x p matrix; and the cluster i, denoted as C_i, is the set of all interest patches that fell into the i^{th} cluster, then we have the following: $P_i = \frac{1}{|C_i|} \sum_{j=1}^{|C_i|} ip_j$, where ip_j is the jth interest patch that was clustered in cluster C_i and $|C_i|$ is the cardinality of i^{th} cluster.

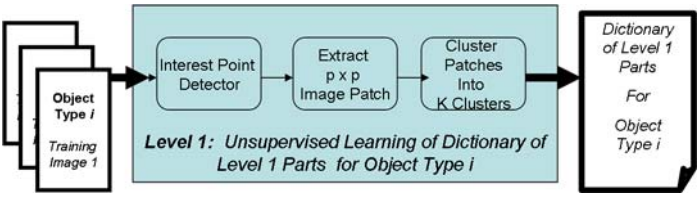

Fig. 3. Level 1 Learning of Parts

The distance metric used for clustering the image patches is the Normalized Grayscale Cross- Correlation, NGCC. If ip_j is the j^{th} interest patch and P_i is the i^{th} part (cluster center), then the NGCC between the interest patch and the cluster center is the following: $NGCC(ip_j, P_i) = \frac{\sigma^2_{ip_j,P_i}}{\sigma_{ip_j}\sigma_{P_i}}$.
Where $\sigma^2_{ip_j,P_i}$ is the covariance between the interest patch and the cluster center, or equivalently, object part; and σ_{ip_j} and σ_{P_i} are the respective standard deviations of the interest patch and the object part.

This Level 1 learning module is extremely simple and fast as it is an unsupervised learning approach to learning the dictionary of parts. However, the number of parts chosen to represent an object type significantly affects the quality of the dictionary of parts. Careful consideration and several trials went into the final value of $K_i = 260$ parts for each object parts-based representation.

Level Two Learning of Parts. In the second level of learning, the object parts $P_1, P_2, \ldots P_{K_i}$ learned for representing object type i increase by one more level in sophistication of their meaning for representing the object type. For Approach 1, this is done by learning the nonparametric two-dimensional spatial frequency distribution for each part for an object types' parts-based representation.

The input of the level 2 learning module for the dictionary of parts for an object type is the set of training images for this particular object type. The output is the set of two-dimensional spatial frequency distribution estimators for each dictionary part of that object type. Thus there are K_i distribution estimators for the object type 1 dictionary of parts. As a result, each level 1 object part, $P_1, P_2, \ldots P_{K_i}$ has a nonparametric spatial distribution attached to it, thus producing a level 2 set of parts to represent object type i.

Classification Using Approach 1 Parts-based Representation. Classification of an object's type is similar to Level 2 learning module. It attempts to capture the level-2 parts that are in the test object and compare it with the parts of each object type learned a priori.

First, an interest point detector is applied to find the interest points. Second, patches of size p x p are extracted, called interest patches. Third, the NGCC distance is computed between each interest patch and each object part of *each* object type. Only those interest patches which yield an NGCC above a threshold pass to the next step. The next step is to see (1) which object type received the most *matches* with its parts and the interest patches, (2) which object type

received *matches* whose interest patches had similar spatial distributions to those of the object parts, and (3) which object type received the highest distance. The object type which satisfies all three criteria is chosen.

2.2 Approach 2 Example Parts-Based Object Recognition Classifier

In this section, a parts-based classifier is presented that takes on the style of Approach 2 of multi-level learning of parts-based object representations. In the second approach to multi-level learning of parts-based representations, the highest level of learning, which is again $L = 2$, involves all possible object types in order to choose parts that better discriminate between object type, as portrayed in Figure 2. This approach is similar to that of [2] except that in our approach a multi-scale-class grouping algorithm was adopted to improve the recognition accuracy.

Level One Learning of Parts. As depicted in Figure 4, the level 1 learning module takes as input training images of a particular object type and outputs a set of level-1 parts that describe that object type. This approach is also unsupervised in the way that it does not have labeled parts already to work with. It discovers the parts itself. However, it does not use clustering to yield the level-1 parts, as in the example for Approach 1. This level 1 module happens to use convolution with several types of filters to create more than one representation for each interest patch extracted from an object in the training image. More specifically, it applie8s 2D convolution with four filters: a delta function, x and y derivatives and a Gaussian, see Figure 5 and the following equation: $v_i(x, y) = [(I * f_i) \otimes P_i] * l_x^T l_y$, where $*$ is the convolution operator, $\otimes$ is the normalized cross correlation operator, $v_i(x, y)$ is the feature vector entry i, f is a filter, P is a patch, and l_x and l_y are the x, y location vectors with respect to the center of the image respectively. The battery of filters f used are depicted in Figure 5. The patches' sizes are selected randomly between 9x9 to 25x25 since it showed better results than using a fixed patch sizes. The location information for the interest patch is recorded as well as the four responses, or level one parts, from the filtering. Location is stored in two Gaussian 1D vectors, where each has an offset equal to the x and y distances, respectively. This distance is computed in a constant time. Note, each level-1 part, P_i, has one location and four responses associated to it.

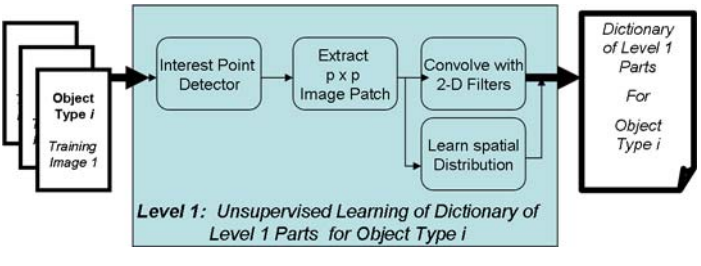

Fig. 4. Level 1 Learning of Parts

Fig. 5. A delta function, x and y derivatives and a Gaussian

Level Two Learning of Parts. In the second level of learning parts, the learning module takes as input the set of level 1 parts learned from *each* object type and outputs a new set of level 2 parts for *each* object type that attempted to maximize inter-class similarities. A joint boosting algorithm is used for multi-class detection and classification, see [2]. This is based on a boosting algorithm where weak learners are sequentially added to form a strong classifier. For each class, a strong learner $H(P_i, c)$ is computed. Where P_i is the level 1 part and c is the object type/class. Each round of boosting, a search is conducted on all the components, f of the level 1 part P_i, for each component, search over all the discrete values of possible thresholds Θ and for each couple f, Θ, find the optimal regression parameters a_S and b_S. Finally, select f, Θ, a_S, b_S that minimizes a cost function. Note, that this level of learning, unlike the previous approach, spans all possible object types/classes in order to choose shared features/parts that attempt to optimize overall object type recognition accuracy.

3 Experimental Results

There were four experiments conducted on a restricted access video sequence of 169 frames called "MOV007_seq1" which used a pan-tilt-zoom camera to record a dynamic field of view including three patrolling Marines that interact closely in distance with each other. Results from one particular video frame is shown in Figure 6. This sequence is part of a collection of clips showing US Marine's outside training exercises as part of a project on behavioral analysis [9]. A dataset including annotated still images from multiple marines poses was used to train the classifiers.

First, the Approach 1 multi-level learned parts-based classifier currently executes in a single scale, so there needs to be a standard person detector first. Thus, the first scenario/experiment is that of running the Felzenszwalb person detector [7] first. Then, the resulting bounding box of the detected person is resampled to a standard size which then inputs to the Approach 1 type posture classifier, which attempts to match the parts of the detected person to the set of parts of each posture learned. As described earlier, the most likely posture

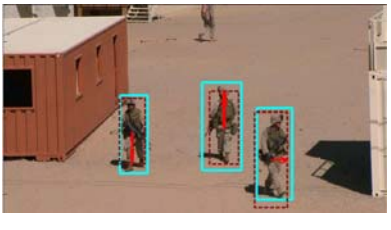

Fig. 6. A video frame with annotated detections/postures

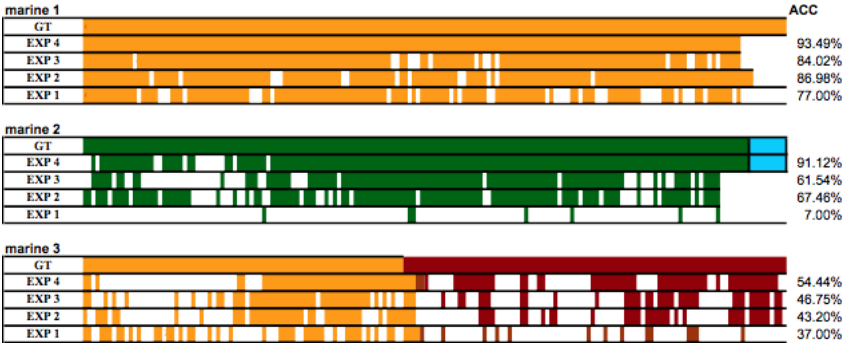

Fig. 7. Temporal analysis and classifier accuracy of all four experiments (GT=ground truth, then Exp. 4, Third row is Exp 3, and fourth row is Exp. 2. Colors: or- 0deg;b-90deg; gr-180deg;r-270deg;blnk-not detected/confused).

is outputted. Note, the detection time is 6-7 sec. The Approach 1 recognition takes 5-6 seconds. The second experiment is similar to the first, except instead of executing the Approach 1 classifier, the Approach 2 classifier, single scale is executed. The detection time is 7-8 seconds and recognition time is 1-2 seconds. The third experiment is similar to the first two, however the Approach 2 example classifier is executed at multi-scale. The detection time in this case is 6 -7 seconds, while the recognition time is 6- 7 sec. Finally, the fourth experiment, the Approach 2 classifier stands on its own and performs both detection and multi-scale posture recognition. This takes 272- 273 seconds.

Figure 7 displays a temporal analysis of the accuracy of the experiments. Going through the video frames, comparing the actual ground truth with the results produced very useful inferences regarding multi-level learning of parts-based recognitions: (1) Since the multi-level learned parts-based classifier using the single-scale Approach 2 learning layout, outperforms the Approach 1 multi-level learned parts-based classifier, it is better it to learn multi-level learned parts which are produced progressively *and* simultaneously for all object parts in a manner to select and share level 2 parts which discriminates between object types' parts; (2) The multi-scale version of the Approach 2 classifier outperforms the single-scale version of the Approach 2 classifier, inferring that resizing the detected persons reduces recognition accuracy.

Finally, for a large period of the video sequence, Marine 2 and Marine 3 are overlapping, so Marine 2 was not detected often because the detector creates only one large detected bounding box for both Marines. Also, when a Marine walking torso 0 degrees has his head turned sideways, the posture recognition classifier gets confused. Finally, Marine 2 was walking 180 degrees (away from camera) most of the time; however Approach 1 classifier classified him mostly as walking 0 degrees (toward camera). (3) From these latter statements, one can infer that with the addition of a head detector and head orientation classifier, the latter three problems would be mitigated. The head detector and head orientation recognition can be thought of as a third level in learning parts for postures,

since it is more sophisticated part than the level 1, and level 2 parts proposed in this paper. In conclusion, future work entails the addition of a head detector with head orientation, a level 3 learned part, to the Approach 2 of parts-based representation of US Marine postures.

References

1. Ross, D.A., Zemel, R.S.: Learning Parts-Based Representations of Data. J. Mach. Learn. Research 7, 2369–2397 (2006)
2. Torralba, A., Murphy, K., Freeman, W.: Sharing visual features for multiclass and multiview object detection. In: IEEE PAMI (2007)
3. Fergus, R., Perona, P., Zisserman, A.: Object recognition by unsupervised scale-invariant learning. In: CVPR 2003 (2003)
4. FeiFei, L., Fergus, R., Perona, P.: OneShot learning of object categories. In: PAMI 2006 (2006)
5. Leibe, B., Leonardis, A., Schiele, B.: Robust Object Detection with Interleaved Categorization and Segmentation. In: IJCV (2007)
6. Wachs, J.P., Goshorn, D., Kölsch, M.: Recognizing Human Postures and Poses in Monocular Still Images. In: IPCV 2009 (2009)
7. Felzenszwalb, P., McAllester, D., Ramanan, D.: A Discriminatively Trained, Multiscale, Deformable Part Model. In: CVPR 2008 (2008)
8. Harris, C., Stephens, M.J.: A combined corner and edge detector. In: AVC (1988)
9. BASE-IT: Behavior Analysis and Synthesis for Intelligent Training, `http://www.movesinstitute.org/base-it/index.html`